30 June 2006 | $10

# Science

Neal Lane is the Malcolm Gillis University Professor and Senior Fellow of the James A. Baker III Institute for Public Policy at Rice University in Houston, Texas. He is a physicist and works on matters of science and technology policy.

# Alarm Bells Should Help Us Refocus

WE'RE HEARING ALARM BELLS THESE DAYS ABOUT SCIENCE IN THE UNITED STATES. ON THE one hand, we've been told that in the global economy of today's "flattened" world, we need to bolster innovation and competitiveness and science and engineering research and education. Earlier this year, when President Bush announced his American Competitiveness Initiative, the future appeared brighter for the physical sciences, math, and engineering (although the National Institutes of Health budget remains flat). But other alarms have sounded that the increases may be at the expense of the disciplines that have historically sought to understand how all this hard work actually helps societies deal with these very issues. Last month, Senator Kay Bailey Hutchinson (R-TX), chair of a Senate panel that oversees the U.S. National Science Foundation (NSF), aggressively argued that the agency should limit its funding for the social sciences and focus on the "hard" sciences. Although the committee stopped short of tying NSF's hands, Congress has yet to make a final decision on whether or not competitiveness is just about technology. Congress should think hard about this.

In the past, investments in science have brought breakthrough technologies, a productive technical workforce and positive trade balance in the high-tech sector, and medical miracles, along with many other tangible benefits. Most Americans believe they are healthier and better off because of the nation's long-standing preeminence in science and technology. Moreover, because other nations are replicating our blueprint for research and higher education with increasing success, competition is growing fierce. So fierce, that our country's present and future position in the world economy is at considerable risk.

All this challenges our political leaders, but it should also challenge the broad scientific community to make sure that our science actually helps provide what most Americans need. Clearly, this requires an aggressive and ambitious program of basic research in the hard sciences, including physics, chemistry, materials science, mathematics and computer science, biology and biomedical science, earth and space sciences, and engineering. But that will not be enough.

Over decades, as our scientific knowledge has become more sophisticated, we have come to recognize how such things as human dynamics and institutional behavior can either enhance or impede the benefits to society of our research achievements. But recognizing that reality is only the first step. We need a much better understanding of how new technical knowledge and tools translate into products, jobs, and wealth; how people learn; how offshoring of jobs, even technical jobs, affects our workforce and quality of life; how increased investment in science and engineering research leads to increased industrial productivity and to better jobs; and how to cope with a host of ever-changing societal problems. These issues are the domain of the social sciences, which also need increased federal support. But that still is not sufficient.

The successful application of new knowledge and breakthrough technologies, which are likely to occur with ever-increasing frequency, will require an entirely new interdisciplinary approach to policy-making: one that operates in an agile problem-solving environment and works effectively at the interface where science and technology meet business and public policy. It must be rooted in a vastly improved understanding of people, organizations, cultures, and nations and be implemented by innovative strategies and new methods of communication. All of this can occur only by engaging the nation's top social scientists, including policy experts, to work in collaboration with scientists and engineers from many fields and diverse institutions on multidisciplinary research efforts that address large but well-defined national and global problems. This will not be easy. It will require qualitative changes in research cultures and in the way federal agencies consider research funding.

Cynics may dismiss these concerns with an abrupt, "We've seen all this before." I believe they are wrong and it would be folly to ignore the alarm bells. Rather, let us use these sometimes shrill warnings to help us refocus and regain the high road for the 21st century for science, the nation, and all of humanity. Albert Einstein eloquently framed this issue for scientists in 1931 at the California Institute of Technology: "Concern for man himself and his fate must always constitute the chief objective of all technological endeavors . . . Never forget this in the midst of your diagrams and equations." Congress, as well as scientists, should remember these words.

**– Neal Lane**

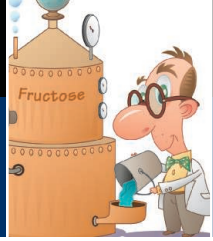
10.1126/science.1131478

ILLUSTRATION: PAT N. LEWIS



Published by AAAS

Empathetic mice?
1860




Plastic from fruit
1861

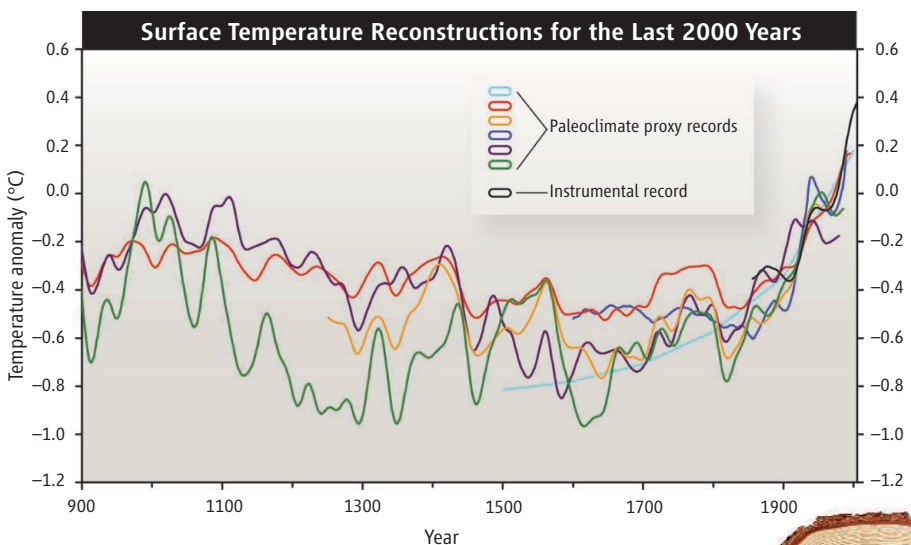
CLIMATE CHANGE

# Yes, It's Been Getting Warmer in Here Since the $CO_2$ Began to Rise

**WASHINGTON, D.C.**—The last decades of the 20th century were most likely warmer than any comparable period in the past 1000 years, a National Research Council (NRC) panel announced* at a press briefing here last week. The expert committee thus confirms the outlines of the near-iconic "hockey stick" temperature curve—a long cooling followed by a sharp warming during the past millennium—that had become a favorite target of greenhouse contrarians. But the committee also says the evidence in parts of the stick is fuzzier than the public and many scientists might have thought.

The hockey stick arose from work published in 1998 and 1999 by statistical climatologist Michael Mann of Pennsylvania State University in State College and two colleagues. They compiled 12 Northern Hemisphere temperature records spanning the past millennium, using climate proxies such as the width of tree rings and the chemical composition of corals. The resulting temperature curve sloped gently downward for most of the millennium (the handle of the hockey stick), then rose sharply into the 20th century (the blade) until it topped the relative warmth of 800 to 1000 years ago. That turnaround suggested that humans played a hand in the recent warming.

After the hockey stick appeared prominently in a 2001 international climate assessment, the critics rushed in. Skeptics said Mann and colleagues had erred badly in their statistical analysis, and some hinted at deliberate distortion.



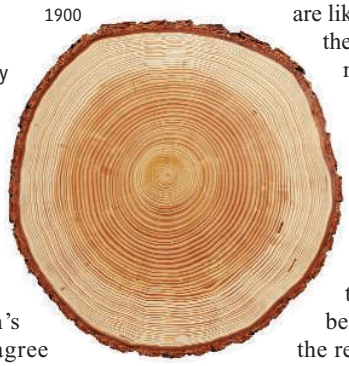

**Warped sticks.** The latest millennial temperature records (produced since the "hockey stick" came out using proxies such as tree rings) may have more squiggles, but they support a recent sharp warming to record high temperatures.

The NRC committee, chaired by meteorologist Gerald North of Texas A&M University in College Station, generally supported Mann's work. "We do roughly agree with the substance of their finding," said North. Mann's group sometimes erred, the committee found. "Some of their choices could have been made better," said statistician and committee member Peter Bloomfield of North Carolina State University, Raleigh, "but it was quite plausible at the time." In any case, the missteps "didn't have a material effect on the final conclusion," he said. And similar studies have followed from a half-dozen other groups, all giving the warm-cool–much warmer pattern.

In addition, none of the three committee members at the press briefing—North, Bloomfield, and paleoclimatologist Kurt Cuffey of the University of California, Berkeley—had found any hint of scientific impropriety. "I certainly did not see anything inappropriate," said North. "Maybe things could have been done better, but after all, it was the first analysis of its kind."

Although the committee generally supported the work Mann led, "there's a disagreement about how sure we are" about some of the study's conclusions, said North. The committee has "high confidence" that the late 20th century was the warmest period of the past 400 years—a time when high-precision proxy records are abundant. That's consistent with the idea that recent warming was in large part human-induced, Cuffey noted. But the committee has "less confidence" in Mann's conclusion that recent temperatures have set a record for the entire millennium. "The committee concluded that Mann and his colleagues underestimated the uncertainty" in the earlier part of the record, said Cuffey, for which records are of lower quality and fewer in number. "In fact, these uncertainties aren't fully quantified," he said.

When pressed, statistician Bloomfield characterized the committee's lesser confidence in the millennial result as "more at the level of 2:1 odds" that Earth is now warmer than it has been in at least 1000 years. The committee has "even less confidence" in Mann *et al.*'s 1999 conclusion that "the 1990s are likely the warmest decade, and 1998 the warmest year, in at least a millennium." "That's plausible," said Cuffey. "We don't know if it's true or not." A year or a decade is just too short an interval for comparison to the older paleotemperature record, he said.



Whether 2:1 odds for a millennial record are good or poor turns out to be in the eyes of the beholder. Long-standing critics saw the report confirming that the hockey stick had not stood up to scrutiny; defenders saw support for key findings. The committee, for its part, stressed that the hockey stick and other records resembling it are not the only evidence of human-induced warming, "and they are not the primary evidence." Cuffey, for one, argued staunchly that the case for anthropogenic global warming is compelling, with or without the hockey stick.

**–RICHARD A. KERR**

CREDITS (TOP TO BOTTOM): ADAPTED FROM NRC; ROYALTY-FREE/CORBIS

* *Surface Temperature Reconstructions for the Last 2,000 Years*, National Research Council, available at fermat.nap.edu/catalog/11676.html


Published by AAAS

AVIAN INFLUENZA

# Journal Letter Spotlights China's Bird Flu Reporting

A bizarre episode surrounding the publication of a letter in this week's *New England Journal of Medicine* (*NEJM*) has again focused attention on China's willingness to share public health information.

The letter details the case of a 24-year-old Chinese man who died in November 2003. SARS was initially suspected, the Chinese authors report, but tests for the SARS virus were negative. Subsequent tests on the patient's stored tissue samples turned up the H5N1 avian influenza virus, the letter states. The death occurred 3 months before China officially reported any H5N1 outbreaks in poultry and 2 years before it officially reported its first human case.

The *NEJM* letter might not have made such a splash were it not for a last-minute attempt to retract it. On 21 June, the day before publication, *NEJM* received e-mails purportedly from corresponding author Wu-Chun Cao at the State Key Laboratory of Pathogens and Biosecurity in Beijing requesting that the article be withdrawn. Because the issue had already been printed, *NEJM* editors sent an e-mail to the journal's media subscribers and posted a note on its Web page advising that the letter had been retracted. Then on 23 June, *NEJM* sent out another announcement saying that Cao had contacted the journal by phone and fax and claimed that he had not sent the e-mails and had not requested that the report be withdrawn. "And so it stands as published in the issue of June 22," reads the e-mail from Jeffrey Drazen, *NEJM* editor-in-chief. The episode prompted speculation about government censorship, but the interpretation remains murky. *Science* could not reach Cao or his co-authors for comment.

Masato Tashiro, director of the World Health Organization (WHO) Collaborative Center for Influenza Surveillance and Research at Japan's National Institute of Infectious Diseases, says the initial suspicions about SARS described in the report are understandable. "Clinically, SARS looks just like avian influenza," he says. Nor would it be surprising if cases of bird flu were missed among the thousands of patients presenting with flu symptoms and pneumonia in China each year, especially in the early days of the outbreak, he says.

Flu experts have long suspected that H5N1 was circulating either undetected or unreported in southern China, probably since the first outbreak of the disease in Hong Kong in 1997. Among other evidence, two members of a Hong Kong family tested positive for H5N1 after a trip to the mainland's Fujian Province in February 2003, and bird flu is suspected in a third member of that family who died in Fujian.

Roy Wadia, a spokesperson for WHO in Beijing, says the timing of events raises questions "about the information-sharing mechanism." Cao's institute comes under the Academy of Military Medical Sciences, and his co-authors are affiliated with the ▶

**Family tragedy.** Surviving members of the extended Ginting family, which lost seven members to bird flu, gather to mourn.

CREDIT: JEAN CHUNG/ONASIA

## Human Transmission But No Pandemic in Indonesia

Bird flu experts meeting in Jakarta last week concluded that a rare instance of human-to-human-to-human transmission had indeed occurred within a large family cluster in Indonesia. But they said there is no sign that the virus is becoming more dangerous and also discounted criticism of the government's handling of the cluster, which occurred in rural northern Sumatra in May.

With 51 human cases and 39 deaths reported so far, Indonesia is the second-hardest hit country after Vietnam. Vietnam brought its bird flu outbreak under control last year, but the number of poultry outbreaks and human cases continues to rise in Indonesia. The Sumatra family cluster is the largest documented to date.

Epidemiologic and genetic sequencing data suggest that a 10-year-old boy contracted the virus from his aunt and then passed it on to his father, concluded the experts, who were convened by the World Health Organization (WHO). (Six blood members of the family have died of H5N1 infection, and it is suspected in a seventh member who was buried before tissue samples were collected.)

Although such localized "second generation" transmission has never been confirmed before, it is not unduly alarming, says Masato Tashiro, director of the WHO Collaborative Center for Influenza Surveillance and Research at Japan's National Institute of Infectious Diseases. The experts found evidence that the virus had mutated slightly as it circulated among family members, but the changes occurred in a genetic region that does not affect transmissibility, he says.

Steven Bjorge, a WHO epidemiologist in Jakarta, defended the country's handling of the cluster, noting that health officials began investigating 2 days after family members appeared at a private clinic. He admits that the country faces an uphill battle in containing poultry outbreaks. Progress is being made tracking outbreaks in poultry with a pilot surveillance scheme that involves local officials and citizens, says Bjorge. But the country will need financial help if it is to extend such a program across Indonesia's 17,000 inhabited islands, which stretch over three time zones. Despite pledges made at an international donor meeting in Beijing in January, Indonesian officials said last week that not one cent had arrived in their country. **–D.N.**

Institute of Microbiology and Epidemiology, a People's Liberation Army hospital, and the Beijing Genomics Institute. Wadia believes that China's Ministry of Health was unaware of this case until news of the *NEJM* paper started circulating just before its publication. WHO has asked the ministry to investigate. "There is a public health significance [to the timely sharing of information] that can't be stressed too strongly," says Wadia.

A prominent Chinese military biologist who asked not to be identified says that Chinese civilian and military researchers often do not share key research results because of fears that findings will be poached. Although he has no direct knowledge of the *NEJM* letter, he speculates that "it is most likely that the H5N1 patient was hospitalized in a military hospital"; otherwise, the military-affiliated research group would never have acquired the tissue samples.

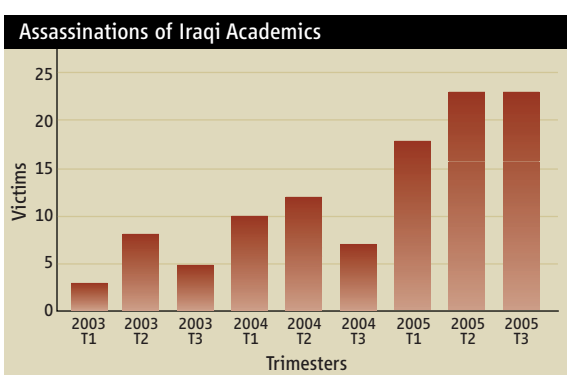
**–DENNIS NORMILE**

With reporting by Gong Yidong and Jia Hepeng in Beijing.

WAR ZONE

# Targeted for Murder, Iraqi Scientists Named on a Hit List

If you want to know how bad it is for scientists in Iraq these days, just ask Nazar Al-Anbaky. In the spring of 2005, a close friend, agronomist Awad Esa, director general of the Ministry of Agriculture's extension division, was gunned down by masked men as he was leaving work. Another colleague, Rafid Abdal Alkareem, head of the animal-welfare board, fled Iraq after surviving two assassination attempts. Faced with persistent threats, the ministry last fall dispersed most personnel around Baghdad. "I wasn't able to do my work. The danger was everywhere," says Al-Anbaky, who was deputy chief of the ministry's plant protection research department. So in March, he too quit Iraq.




**Mounting toll.** The murder rate of Iraqi academics has risen steadily since the April 2003 invasion.

CREDIT (TOP TO BOTTOM): UC SANTA CRUZ; ADAPTED FROM *PLIGHT OF IRAQI ACADEMICS*, I. JALILI, MAY 2006

For months, Iraqi academics have denounced what they view as an unspoken campaign to cripple the country's intellectual elite (*Science*, 30 September 2005, p. 2156). Now they face an overt new threat. An unidentified group is circulating a hit list of 461 Iraqi intellectuals. The existence of leaflets calling for the assassination of named individuals was reported by the newspaper *Az-Zaman* last month; *Science* has obtained a copy of the list, verified by several Iraqi scientists as authentic. Last week, rectors of six universities in Spain issued a statement warning of "a very grave outrage against the cultural and scientific development" of Iraq and urging authorities to investigate "the killing campaign."

For Iraq's beleaguered scientists, the hit list aggravates a desperate situation. Since the U.S.–led coalition invaded in April 2003, at least 188 Iraqi academics have been slain, according to a tally by the Spanish Campaign Against Occupation and for the Sovereignty of Iraq, based in Madrid. Over the past 3 years, the pace has increased (see graph). In that period, some 220 doctors have been killed and more than 1000 have left Iraq, the health ministry reported last February. Hundreds of scientists have fled the country. "This brain drain will adversely affect Iraq's development for years to come," says Jafar Jafar, head of Iraq's nuclear program under Saddam Hussein. Jafar, general manager of Uruk Engineering Services in Dubai, says he has helped "many friends and acquaintances" find jobs elsewhere.

The killers are largely unknown. Some murders are sectarian: Sunni militias targeting Shiite academics and vice versa. Overall, however, the assassinations "do not follow any religious or sectarian pattern," says Ismail Jalili, an ophthalmic surgeon who presented an in-depth analysis at a conference in Madrid last April.

In some cases, money is a motive. One recent victim was Ali Hassan Mahawish, dean of engineering at Al-Mustansiriya University in Baghdad, who told *Science* last September how several professors in his department had gone overseas on sabbatical, depleting the faculty. He was seized by gunmen in March. ▶

## Denice Denton (1959–2006)

Denice Denton, chancellor of the University of California (UC), Santa Cruz, and a champion of diversity in science and engineering, jumped to her death from a San Francisco apartment building on 24 June.

An electrical engineer, Denton became the first female dean of engineering at a major research university when she came to the University of Washington (UW), Seattle, in 1996. At UW, she established programs to bring more women and minorities into engineering and introduced policies to enable female faculty members to balance work and family. "She had the ability to make everyone feel included," says Eve Riskin, an electrical engineering professor at UW.

In February 2005, she moved to UC Santa Cruz, where she was criticized for helping her partner, materials scientist Gretchen Kalonji, get a UC administrative job and for the $600,000 spent on renovations to her campus home.

Last fall, Denton expressed frustration about her job during a meeting on women and science at the U.S. National Academies. "It's lonely at the top. No one has on their list of things to do, 'Be nice to the dean or the provost today. [Ask yourself] what can I do to support them in their endeavors for social justice?' " A colleague said Denton confided on 3 June that she was "very demoralized" and "didn't know how much more she could take."

"Denice was an accomplished and passionate scholar whose life and work demonstrated a deep commitment to public service and to improving opportunity for the disadvantaged and underrepresented," says UC President Robert Dynes. An interim chancellor is expected to be appointed soon.

**–YUDHIJIT BHATTACHARJEE**

## A Taste for Variety

**MELBOURNE, AUSTRALIA**—The Human Variome Project—a planned database of all variant forms of human genes and their phenotypes—got off the ground here last week. Fifty-five international experts met to lay out a framework and name geneticist Richard Cotton their chief. Cotton says the project, which aims to cull global mutations data from databases and medical records to understand disease, needs $60 million.

**–ELIZABETH FINKEL**



Published by AAAS

"The ransom was paid, but his family got a dead body," says a colleague in Baghdad who asked to remain anonymous. The latest drama involves petroleum scientist Muthna Al-Badery, a top official in the Oil Ministry, who was kidnapped earlier this month. "Bargaining is still continuing for his life," the Baghdad scientist says.

The hit list includes scientists, university officials, engineers, doctors, and journalists in Baghdad and other cities. "The list is part of an organized, foreign-backed campaign to terrorize Iraqi brains," an official with the Iraqi Writers Union told *Az-Zaman*. No one contacted by *Science* knows who issued the list. One prominent scientist with ties to Iraq's intelligence community says that Iraqi investigators are probing claims that Iranian intelligence agents are involved. The U.S. Embassy was not aware of the list, says spokesperson Dennis Culkin.

One thing is certain: The campaign has cast a pall over Iraqi academia. Says one engineering professor who is sticking it out in Baghdad, "We carry our coffin every day we go to work."

**–RICHARD STONE**

ETHICS

# Blocking a Book, Dutch University Rekindles Furor Over Nobelist Debye

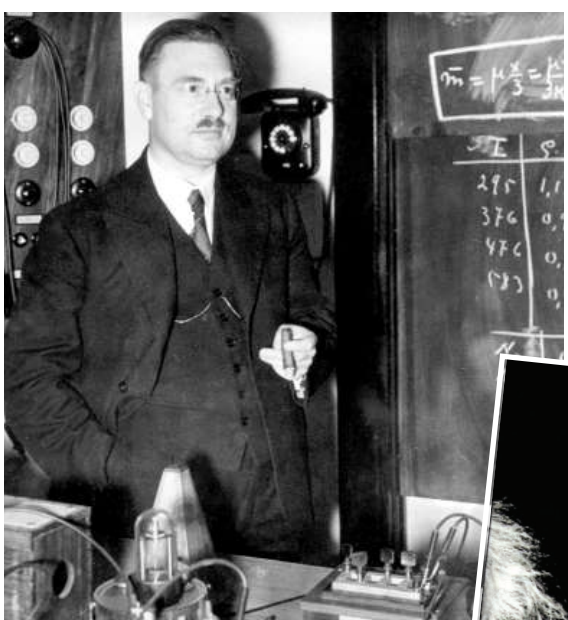
A controversy about the alleged Nazi sympathies of Dutch chemistry Nobel laureate Peter Debye has escalated. Utrecht University last week halted publication of a pro-Debye book by an employee and ordered staff not to discuss the issue with the press. The move follows a university decision last February to strip Debye's name from its institute for nanomaterials.

A science historian, meanwhile, has spoken out in Debye's defense, as has another Dutch Nobel laureate, Martinus Veltman. Cornell University, where Debye was a professor from 1940 until his death in 1966, has concluded from its own 3-month investigation that there's no reason to distance itself from him, as has the American Chemical Society (ACS).

The flap erupted after the publication of a harsh view of Debye—a physical chemist who led the Kaiser Wilhelm Institute for Physics in Berlin from 1935 to 1939—in *Einstein in Nederland: Een intellectuele biografie* by Berlin-based journalist and science historian Sybe Rispens. One chapter, excerpted in a weekly magazine, documented that Debye, as president of the German Physical Society (DPG), asked Jewish members to resign in a 1938 letter signed "Heil Hitler!" It also claimed that Debye stayed in touch with German authorities while at Cornell, even offering to return to Berlin in June 1941.

In a brief statement issued on 16 February, Utrecht University's board said it would rename the Debye Institute, and Maastricht University said it would no longer award the Peter Debye Prize (*Science*, 3 March, p. 1239). Gijs van Ginkel, managing director of the "former Debye Institute," as it now calls itself, responded by writing a book containing an analysis of historical documents, his view of the affair, and a sharp attack on Rispens.

But the university has halted its publication. Van Ginkel referred questions to university spokesperson Ludo Koks, who denies that academic freedom is at stake; Koks says Van Ginkel had broken an agreement not to include personal comments in the publication. Koks confirms that institute staffers have been ordered not to talk to the press to "streamline communications."




**Counter-counterattack.** Utrecht University has halted the publication of a book that countered allegations about Peter Debye published in an earlier book (*inset*).

Mark Walker, a historian at Union College in Schenectady, New York, who studies science and technology in the Nazi era, says that although Debye "didn't show civic courage, … all the evidence is that he was not a Nazi sympathizer." For example, the DPG purged its Jewish members much later than most other scientific societies did, and without any enthusiasm whatsoever, he says. Signing official letters with "Heil Hitler!" was nothing unusual, even among those openly opposed to the regime. That Debye tried to keep communication channels to Germany open while at Cornell is also "absolutely reasonable," Walker says, because his daughter still lived there.

Walker recently gave a lecture about Debye at Cornell, where the affair "was something we just couldn't ignore," says Héctor Abruña, chair of the Department of Chemistry and Chemical Biology. "Debye has had such a huge influence here." In a 1000-word letter submitted for publication to *Chemical and Engineering News*, Abruña says a review shows that removing Debye's name from a professorship and a lecture series would be "unwarranted." Banning books is "not what universities should be about," Abruña adds.

ACS sees "no compelling reason to do anything" about its Peter Debye Award in Physical Chemistry either, says Gordon McCarty, chair of ACS's Committee on Grants and Awards; DuPont, the awards sponsor, is "quite comfortable" with that stance, he adds.

Rispens says he opposes silencing different views on Debye and would welcome a study that went beyond his own focus on Albert Einstein's circle. But the affair has cost Rispens the support of one enthusiastic fan: Veltman, who, in a foreword to Rispens's book, praised it as "a nugget of gold." In a 5 May open letter to Debye Institute staff, Veltman says he took Rispens's assertions "at face value" at the time but now realizes "they should be assigned to the realm of fables." The foreword will not appear in new editions or translations of the book, Veltman continued; the two universities "should admit their error, revoke their decision, and forget the matter," he says.

**–MARTIN ENSERINK**

CREDIT (LEFT): AP PHOTO

DIVERSITY

# Report Urges National Academies To Improve Status of Women

Johanna (Anneke) Levelt Sengers stands at the top of her profession but confesses that "it can be a little lonely" as one of only two women in the 82-member engineering sciences section of the U.S. National Academy of Sciences (NAS). A scientist emeritus at the National Institute of Standards and Technology, she belongs to both NAS and its partner, the National Academy of Engineering, where she's one of seven women within the 173-member chemical engineering section. So in late 2004, when she was asked to co-chair an international panel on women in science with Manju Sharma of India, they decided to examine not just women's place in society but also their status within the 90 national academies that had requested the report.


**Against the odds.** Levelt Sengers helped write a report on women for an international council of national academies whose 15-member board (shown *below* in January 2006) is all male.

The report, posted last week by the InterAcademy Council (IAC) (www.interacademycouncil.net), offers a refreshingly candid assessment of the problems facing women trying to enter and move up in the world of science and engineering. Although it strikes familiar chords about the need to remove barriers and increase opportunities for girls and women, it sings a new tune in commanding the national academies themselves to "first put their own houses in order." In addition to choosing more women as members and leaders of their organizations, each national academy should form a standing committee on diversity to gather and discuss gender-related data, it says.

"Wow. This is far more hard-hitting and to the point than I had expected," says Donna Dean, president of the Association for Women in Science in Washington, D.C., and a former senior administrator at the National Institutes of Health, who is now at the Washington, D.C., science-lobbying firm of Lewis-Burke Associates. "It tells the various academies to stop pontificating about the right thing to do and start showing it in how they operate."

The report was funded in part by a $50,000 grant from L'Oreal. Since 1998, the France-based cosmetics company has honored outstanding women scientists around the world—including five of the eight women on the 10-person IAC panel. Jennifer Campbell, who heads the company's philanthropic efforts, says she would like to see across-the-board parity for women in science. But Levelt Sengers says she thinks that "a reasonable goal would be no major disparity between the percentage of Ph.D. degrees awarded to women in a particular field and the percentage elected in that field." Most academies are a far cry from reaching even that level; the 2% figure for NAS women in chemical engineering, for example, pales beside the 14% of U.S. Ph.D.s awarded in the 1980s, much less the 22% awarded in the 1990s.

NAS President Ralph Cicerone says that there's "no magic bullet" for adding women to the academy's ranks but that NAS is trying to increase their chances of gaining the type of recognition—through service on academy panels, keynote speeches, and major scientific awards—that traditionally leads to NAS membership. NAS has no plans "to collapse its activities into one committee on gender issues," he says, adding that the challenge calls for "a sustained effort … along the entire pipeline."

Levelt Sengers says that each academy must come up with individual remedies, which she hopes will be discussed during the council's next meeting in December in Cairo, Egypt. Dean suggests a radical approach to staffing the academies, many of which operate extensive networks of institutes and laboratories. "What about strategic buyouts to senior managers, like companies do?" she asks. "It wouldn't be easy or welcome. But business as usual just won't get you there."

–JEFFREY MERVIS

CREDITS (TOP TO BOTTOM): B. EYMANN, COURTESY OF THE FRENCH ACADEMY OF SCIENCES; Y. O'BANNON, COURTESY OF THE INTERACADEMY COUNCIL

## Discovery Carries Heavy Load

A successful trip for space shuttle Discovery, set for launch this weekend, will boost prospects for repairing the Hubble Space Telescope and completing the half-built international space station. Despite concerns by safety officers, NASA approved the second shuttle mission since February 2003, when Columbia disintegrated upon return. Even if all goes smoothly, the agency said last week, it plans to scale back U.S. research aboard the station by eliminating a centrifuge and other scientific equipment.

–ANDREW LAWLER

## Get Your Shots

Thirty British doctors this week called for responsible media coverage amid published doubts about the measles, mumps, and rubella (MMR) vaccine. Concerns over MMR have accompanied a decline in the number of vaccinated U.K. children from 93% in 1995 to 83% in 2005, and the doctors cite a "dramatic" rise of measles this year as well as a fatality, the first in 14 years. To blame, they say, is a 1998 *Lancet* paper linking the jab to autism by Andrew Wakefield, who was charged with misconduct this month by the U.K.'s General Medical Council. "Illness or death" could befall unimmunized children, the signatories warn.

–LAURA BLACKBURN

## Patently Obvious? Ask The Supremes

Is the U.S. government granting patents for inventions that are obvious? This week, the U.S. Supreme Court accepted a case that biotech attorneys say could make new patents harder to obtain. *KSR International* v. *Teleflex* involves a dispute over a gas pedal, but the case has "tremendous implications for biotech," says Hal Wegner of Foley & Lardner LLP in Washington, D.C. That's because the court could toughen a standard used to determine the validity of an application that combines elements of published ideas or patents. The standard is whether the published work includes a specific "suggestion" to combine existing parts. A 2004 National Academies panel called for "a stricter standard" to improve biotech patent quality. Arguments are set for autumn in the case, which is expected to pit the software industry against the biotech and pharma sectors.

In other patent news, last week the justices decided not to act on a case they had heard involving whether scientific information can be patented (*Science*, 17 February, p. 946).

–ELI KINTISCH



Published by AAAS

PALEOCLIMATOLOGY

# Atlantic Mud Shows How Melting Ice Triggered an Ancient Chill

Eighty-two hundred years ago, a chill swept around the Northern Hemisphere, the last, feeble gasp of the mighty 100,000-year ice age that preceded it. Geologists looking for a cause had glimpsed evidence that a vast outpouring of glacial meltwater had gushed into the Atlantic Ocean less than a year after an ice dam busted. That meltwater outburst, seven times as voluminous as all five present-day Great Lakes, had come suspiciously close to the 8200-year, or "8K," chill. But paleoceanographers couldn't answer the big question: How could it have affected deep-sea currents believed to play a pivotal role in controlling climate?

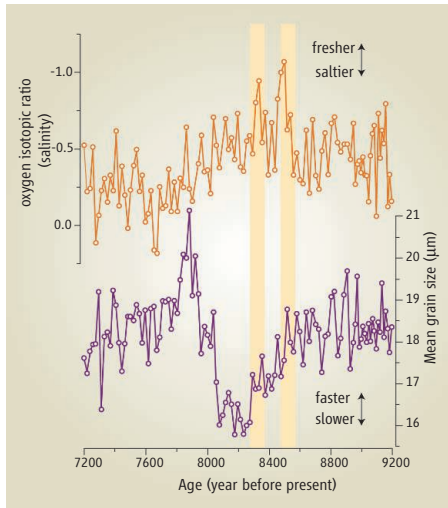
Now, paleoceanographers report on page 1929 that they have found a single ocean sediment core that preserves the sought-for link. An ocean current, the so-called conveyor, carries climate-moderating heat into the far north, where it sinks to the bottom and heads south. The core recorded both a gush of freshwater far out into the North Atlantic and the nearly simultaneous slowing of that conveyor. "I'm a believer" in the meltwater-conveyor-climate link, says paleoceanographer Nicholas McCave of Cambridge University in the U.K., who is not an author of the paper.

The core came from south of Iceland on the Gardar Drift, where muddy sediment collects 10 to 20 times faster than is usual in the deep sea. With more sediment per year to work with, paleoceanographers Christopher Ellison and Mark Chapman of the University of East Anglia and Ian Hall of Cardiff University, both in the U.K., could sample smaller bits of time and thus read a more detailed history. To gauge the changing temperature of surface waters, they measured the abundance of a cold-loving plankton species. They calculated the salinity of the seawater from the ratios of oxygen isotopes in the plankton's shells after adjusting for the effect of temperature. Finally, they inferred the speed of ancient currents from the varying size of silt particles in the mud. The more abundant the larger particles were, the faster the current was moving across the bottom.

By reading the history of both surface and bottom waters in a single core, the U.K. researchers nailed down the order of events more than 8000 years ago. They found that the cold meltwater came in two pulses, the first about 8490 years ago, the second 200 years later. Late in the first freshening, the returning conveyor current sweeping southward along the Gardar Drift began to slow. Apparently, the fresher, more buoyant surface waters slowed the sinking of the conveyor's waters into the deep sea at the current's turnaround point. The second freshening repeated the pattern and further slowed the conveyor.

The Gardar results are "a strong confirmation that this was a freshwater event," says geochemist Wallace Broecker of Lamont-Doherty Earth Observatory in Palisades, New York. That's reassuring, Broecker says, because it helps clear up a vexing puzzle about an earlier, even more drastic cooling: the 11,000-year-old Younger Dryas cold spell, also supposedly due to a glacial outburst. Broecker and geologist colleagues reported recently that they couldn't find the route of the meltwater on land. That failure raised the troubling possibility that glacial meltwater had nothing to do with the Younger Dryas, the 8K event, or other abrupt coolings of the past 15,000 years. Now it looks as if theorists were on the right track after all.

Next, researchers need to figure out why the far smaller meltwater release of the Younger Dryas triggered a cooling so much greater than the 8K's. "The 8K was the biggest flood," says glaciologist Richard Alley of Pennsylvania State University in State College, "but it's the one that didn't stick," climatically speaking. Had the climate system by then developed some protective property that warded off the abrupt chill? If so, will the present climate be able to do the same as the greenhouse sends more freshwater—rain and Greenland meltwater—into the North Atlantic? **–RICHARD A. KERR**




**In sync.** When North Atlantic surface waters freshened (*top*), bottom currents slowed.

ANIMAL BEHAVIOR

# Signs of Empathy Seen in Mice

Empathy is one of the nobler human attributes, which may explain why we're often reluctant to ascribe it to other animals. A debate has simmered for years about whether chimps display empathy, for example. Now on page 1967, scientists argue that even lowly mice have a rudimentary form of it.

The research team, led by Jeffrey Mogil at McGill University in Montreal, Canada, reports that mice become more sensitive to pain when they see a familiar mouse in pain. That probably doesn't qualify as empathy as the word is understood in everyday conversation, says ethologist Frans de Waal of Emory University in Atlanta, Georgia. Still, de Waal and others say it does suggest that mice have some ability to sense what their fellow rodents are experiencing. "They're in tune with each other," de Waal says.

**Commiserating mice.** Observing a cagemate can influence pain sensitivity in mice.

In one experiment, Mogil and colleagues injected mice in the belly with a weak acetic acid solution. Solitary mice react to the injection with a stereotyped writhing behavior, stretching repeatedly and extending their back legs. Mice spent more time writhing when the researchers placed them in a Plexiglas cylinder with a cagemate—a mouse they'd lived with for at least a week—given the same injection at the same time. When the researchers paired mice who'd never met, however, no significant increase in writhing occurred.

The researchers also injected cagemates with formalin in one paw, causing them to lick the presumably painful area. In some cases, both mice received the same concentration of formalin, either a low dose or a high dose. Not surprisingly, pairs of mice given the high dose spent more time licking their paws than did pairs given the low dose. But when a mouse given the low dose was paired with a cagemate given the high dose, it licked more, on average, than if it had been paired with another low-dose mouse. More importantly, the high-dose mouse licked less, on average, than did a high-dose mouse paired with a cagemate that also got a high dose. Observation, it seems, can reduce pain behavior as well as enhance it, Mogil says. ▶

CREDITS (TOP TO BOTTOM): ADAPTED FROM ELLISON *ET AL.*, *SCIENCE*; MONA LISA CHANDA



Published by AAAS

Finally, the team repeated the acetic acid experiment and incorporated a different gauge of pain sensitivity, measuring the time it took for a mouse to withdraw its paw from a hot spot on the floor of the test cylinder. When observing a cagemate writhing from an acetic acid injection, mice withdraw their feet from the heat more quickly—even if they'd received no injection themselves. That's the most important experiment, Mogil says, because it indicates that the mice aren't simply imitating what they see the other mouse doing. "It suggests the pain system is being sensitized in a general manner" by seeing a cagemate in pain.

Is that empathy? It depends on whom you ask, says Tania Singer, a cognitive neuroscientist at the University of Zurich in Switzerland who has studied pain and empathy in people (*Science*, 20 February 2004, p. 1121). "Philosophers would argue you can only have empathy if you have consciousness," she explains. "Psychologists would want to see evidence of altruistic behavior and altruistic motivation." Mice probably don't meet those criteria, she says.

And not everyone agrees that Mogil's experiments actually address the issue of empathy in the first place. Writhing and paw licking are reflexive behaviors mediated by the spinal cord, notes Charles Vierck, a neurobiologist at the University of Florida, Gainesville. "So what we have here is modulation of a reflex response during observation … of the reflex responses of other animals." And that, says Vierck, is nothing new.

Still, Singer and others, including Mogil, interpret the findings as evidence that mice have "emotional contagion," a primitive kind of empathy. "Emotional contagion means one baby starts crying and all the babies start crying," explains Peggy Mason, a neurobiologist at the University of Chicago who studies pain. Unlike higher forms of empathy, it doesn't require understanding what others are experiencing. "The second baby doesn't have to realize that the first baby is upset because it has a dirty diaper," notes Mason.

Many researchers see emotional contagion as a steppingstone toward the more sophisticated kind of empathy that evolved in humans. "To imagine that empathy just started de novo in primates seems biologically implausible," says Mason. **–GREG MILLER**

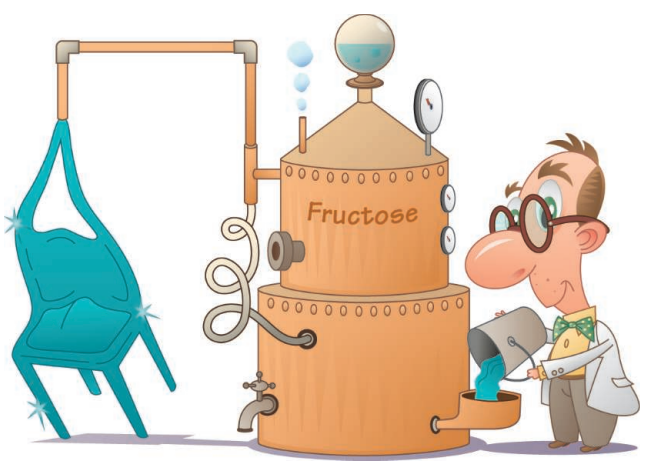
CHEMISTRY

# Sugary Recipe Boosts Grow-Your-Own Plastics

Motorists aren't alone in feeling the pain of rising oil prices. Some commodity chemicals, such as the polypropylene that's found in everything from textiles to dashboards, have tripled in price in the past few years. That surge has spurred new interest in the once-sleepy field of converting crops and other renewable feedstocks into commodity chemicals. Chemical companies have made progress in recent years. But up to now, it has been difficult and costly to make the kinds of compounds that serve as starting materials for most oil-derived chemicals. Now, work by researchers in the United States may give plant-derived chemicals a new push.

On page 1933, University of Wisconsin, Madison, chemical engineer James Dumesic and colleagues report a new process for turning fructose, the sugar in fruit, into a compound called 5-hydroxymethylfurfural (HMF), which can replace key petroleum-derived chemical building blocks. Unlike previous schemes for turning sugar into HMF, the new process is efficient, easy, and potentially low cost. "It looks real good to me," says Thomas Zawodzinski, a chemical engineer at Case Western Reserve University in Cleveland, Ohio. "This is the direction things need to go in."

Because of its current high cost, HMF isn't produced commercially in large volumes. But it is easily converted into other compounds, such as one abbreviated FDCA that can serve as the starting material for commodity chemicals such as polyesters. On paper, converting fructose, a small, ring-shaped sugar, to HMF is simple. It requires stripping off what amounts to three water molecules. Researchers have developed numerous acid-based catalysts that do that quickly. Once formed, however, HMF readily reacts with fructose, other intermediate reactants in the mix, and even itself to form a chemical zoo of unwanted byproducts, which sharply limits the amount of HMF that's recovered at the end. In hopes of solving that problem, researchers have tried adding an organic solvent that sits atop the water like oil in salad dressing. HMF's affinity for the solvent spirits it out of the aqueous phase, during which the unwanted side reactions take place. Unfortunately, separating HMF from the solvents proved difficult. Researchers had to boil away the solvents at very high temperatures.




**Sweet prospects.** Sugars from fruits and grains could replace petroleum as the basis for commodity chemicals such as plastics, if a new process pans out.

To improve the process, Dumesic and his students Yuriy Román-Leshkov and Juben Chheda had to solve these and other problems simultaneously. They did so by adding a series of different compounds both during the aqueous phase, in which catalysts convert fructose into HMF, and to the solvent. In the aqueous phase, the additives—abbreviated DMSO and PVP—suppressed side reactions, thereby encouraging HMF production. Unfortunately, they also increased HMF's solubility in water, making it harder for the solvent, known as MIBK, to remove HMF from the aqueous phase before it could react further. The Wisconsin team overcame that obstacle by spiking the MIBK with a dash of a compound called 2-butanol, which increased HMF's affinity for the solvent. Finally, because MIBK has a low boiling point, the Wisconsin team could easily evaporate it along with the 2-butanol, recover the HMF, and return the solvents to the reactor.

The changes doubled the percentage of fructose that gets converted into HMF, to 85%. With that boost and related improvements, "now you can make some pretty compelling arguments" for producing HMF commercially, says Todd Werpy, an expert on producing bio-derived chemicals at Pacific Northwest National Laboratory in Richland, Washington. Producing commodity chemicals from renewable feedstocks "is really in its infancy," Werpy says. But with top research groups now training their sights on the problem, he adds, "renewables could make a major contribution to the chemical needs in the United States."

**–ROBERT F. SERVICE**

ILLUSTRATION: TERRY E. SMITH



Published by AAAS

**A global initiative to knock out every mouse gene struggles to get its act together**

# A Mouse for Every Gene

**IN ADRIANO AGUZZI'S EXPERIENCE,** getting hold of a new mouse strain can be nothing but trouble. A neuropathologist at the University Hospital of Zurich in Switzerland, he is one of thousands of researchers who study mutant mice for clues to what particular genes do. "Once I requested a mouse, and the guy wanted everyone from himself to his grandmother to be a co-author on everything we published with that mouse," says Aguzzi. "It was like scientific prostitution." Another time, he says, a researcher promised him a mouse but took more than a year to deliver: "[The investigator] should have just said his cat ate it; it would have saved us a lot of trouble."

Most mouse researchers can tell similar horror stories. But help is on the way. Several large-scale projects plan to disable every gene in the mouse genome and make the resulting mice readily available to the research public. In January, Europe and Canada embarked on ambitious efforts that together will produce more than 30,000 knockouts. And this summer, the U.S. National Institutes of Health (NIH) will announce the Knockout Mouse Project (KOMP), which will add another 10,000 to the list. China, too, is gearing up to make 100,000 mutants, with the goal of making 20,000 lines of mice, each with a different gene knocked out. (see sidebar, p. 1864). All told, these efforts will cost almost $100 million. Although separate entities, "the plan is to have every center work together, much like [what] was done with the Human Genome Project," says Allan Bradley, director of the Wellcome Trust Sanger Institute in Cambridge, U.K., which is part of the European effort.

Indeed, overall, the knockout effort is arguably the largest international biological research endeavor since the Human Genome Project. And it is the next major step in figuring out what makes us tick. The human and mouse genome projects each identified some 25,000 genes, most quite similar between the two species. But researchers have no idea what more than half of these genes do. Because the mouse is so amenable to genetic manipulation, and so well studied, mass-produced mutant mice offer a window into these unknown genes. "The Human Genome Project wasn't done just to get the sequence," says Christopher Austin, director of the NIH Chemical Genomics Center and KOMP's founding father. "It was a prerequisite for figuring out what our genes do."

How the individual mass-knockout projects will work together is still being ironed out. Each project is embarking on a different—and not necessarily compatible—approach to making its mutant mice, and the logistics of keeping track of all the mutants made are daunting. In addition, each effort will need to work out an efficient way to catalog and distribute the mice it creates. They will also have to deal with intellectual-property claims when one of the new mutants turns out to be a previously patented mouse strain. "The mouse project could open up huge areas of science, just like the Human Genome Project did," says Marina Picciotto, a molecular neurobiologist at Yale University, "but there are likely to be hiccups along the way."

**Holy Grail?** Marina Picciotto would love to find a mouse that caves to peer pressure, but chances are it's hidden away or hasn't been made yet.

Although Picciotto and most of her colleagues are optimistic about mass-produced knockouts, some wonder whether the efforts are the best use of public resources. Knocking out genes is really just the beginning. Those tens of thousands of mutant mice won't do many researchers much good until the behavior, morphology, and physiology of these knockouts have been described. Characterizing each mouse will not be easy. "You can knock out every gene, but if you don't have assays to evaluate them, it's hard to figure out what the

CREDITS (TOP TO BOTTOM): ED MATHIS/BMC, HEALTH SCIENCES CENTRE,WINNIPEG; TERRY DAGRADI/MED MEDIA GROUP, YALE SCHOOL OF MEDICINE



Published by AAAS

**Buyer beware.** Deactivating the same gene in Black 6 (*left*) and 129 mice may yield widely different phenotypes.

gene is doing," says Marnie Halpern, a zebrafish geneticist at the Carnegie Institution of Washington in Baltimore, Maryland.

### Hiding out

As a group, the knockout projects are trying to create something akin to the international superstore IKEA, where, in a single trip, customers can buy a houseful of easy-to-assemble furniture at reasonable prices. In this case, however, researchers wouldn't even have to make a trip to the store. Ideally, they would simply go to a central database and click their own computer mouse to order the knockout mouse of their choice. Within weeks, frozen embryos would arrive at their door. Like IKEA, some assembly would be required: turning those frozen embryos into live mice. But that requirement is minimal compared to the tens of thousands of dollars and a year or more of work involved in creating an average knockout mouse.

Such a resource would be a far cry from today's mouse trade, which is more like buying furniture from neighbors. Selection is limited, quality varies, and some items just aren't for sale. Part of the problem, says Francis Collins, director of NIH's National Human Genome Research Institute in Bethesda, Maryland, is that until recently, researchers often didn't know what the lab down the street—let alone one in another country—was doing. Investigators aren't required to place their mice in public repositories, and some never write up knockouts they don't find useful.

To remedy this situation, NIH went on a mouse hunt. It started its inquiry at the Jackson Laboratory (JAX) in Bar Harbor, Maine. JAX stores more than 800 varieties of mutants and maintains a database of every published mouse knockout. Then NIH went door-to-door, publishing a request asking investigators go to a JAX Web site and list any knockouts they had created and were willing to share with the research public.

The findings were dispiriting. All told, the mouse community had knocked out about 11,000 genes, but many labs were repeating work done elsewhere. More than 700 knockouts had been created three times or more; in one case, a single mouse had been duplicated 11 times. And of the 4000 unique knockouts that have been published, more than 3000 are not in public repositories, meaning most are either unknown or unavailable to the wider community. "It's embarrassing," says Collins. "A graduate student shouldn't spend a year making a knockout that's already been made. It's not a good use of resources."

Yale's Picciotto is a case in point. As a researcher who studies the genetics of addiction, she would love to find a mouse that caves to peer pressure. So far, she's managed to make a few handy knockouts. Some shun nicotine; others dig opiates. One even seems to be operating on a natural antidepressant. But for a complete picture of the mouse social psyche, Picciotto

CREDIT: LEXICON GENETICS INCORPORATED


**Out cold.** Lexicon is making thousands of mouse knockouts in embryonic stem cells. These frozen lines will become part of the TIGM resource.

## NIH Knocks Out Key Mouse House

When the Texas Institute for Genomic Medicine (TIGM) applied to be part of a new $50 million U.S. National Institutes of Health (NIH) program to knock out as many mouse genes as possible, it seemed to be a shoo-in. Thanks to a partnership with Lexicon Genetics in The Woodlands, Texas, TIGM already has in its freezers knockouts for nearly a third of all mouse genes—twice what global knockout projects have achieved so far (see main text). "Taking us on would have made it easy for [NIH] to fulfill its mission," says TIGM President Richard Finnell.

Instead, he says, NIH has rejected his institute's application, potentially forcing NIH's Knockout Mouse Project (KOMP) to start from scratch and positioning TIGM as a possible competitor. NIH won't comment on the move until it announces the winners of the competition later this summer, but some in the mouse community feel that Lexicon's reputation for tough intellectual property (IP) restrictions may have hurt TIGM's chances. Finnell insists, however, that TIGM will place no IP restrictions on its knockouts.

Founded as a nonprofit organization last summer with a $50 million award from the Texas Enterprise Fund—a $295 million pot set up by the state to create jobs—TIGM's mission is essentially identical to that of the global knockout effort: Establish a massive mouse-mutant resource in embryonic stem cells and distribute these lines to academic scientists at cost. But while the global program's players are just beginning to churn out knockouts, TIGM, which is based in Houston and College Station, has lept ahead.

It has used $30 million of its $50 million to purchase Lexicon's growing library of knockouts in the coveted Black 6 strain of mice; starting this month, researchers can order any of 7500 unique knockouts—representing about a third of the mouse genome—and they'll have access to knockouts covering more than two-thirds of the genome by late 2007, says Finnell.

Becoming part of KOMP would not only have helped NIH achieve its goals more quickly and cheaply, says Finnell, but it would have also made TIGM's mouse lines more economical for researchers. Without NIH support, TIGM will still be supplying knockouts years before KOMP, says Finnell, although researchers may have to pay more for them. (Pricing details are still being worked out.)

Lexicon CEO Arthur Sands is confounded by NIH's decision. "It just doesn't make sense," he says. "[Our] resource is already on the ground." Neither Sands nor Finnell would speculate on why NIH decided not to collaborate with the institute. And outside scientists were hesitant to speak on the record. But some researchers *Science* spoke to said IP restrictions Lexicon has imposed in the past—such as requiring labs and universities to sign away certain rights related to discoveries made using its mice—have been problematic. Under the TIGM deal, however, those restrictions are lifted, says Finnell, "so that wouldn't have been an issue."

Others say NIH is interested in more cutting-edge science than Lexicon is using to make its lines. Ideally, for example, KOMP centers will use gene-specific targeting technology in addition to random gene-trapping technology. According to Finnell, Lexicon's library is being made almost exclusively by means of gene trapping (see figures, p. 1865), although he says that—with NIH funding—TIGM would have tried to complete the remaining third of the resource using gene targeting.

Despite the NIH setback, TIGM is planning to make its mark in the mouse world. "It will cost more now, but we're going to get these lines out to researchers," says Finnell. "When people think about knockout mice, they'll think about TIGM." **–D.G.**


Published by AAAS

## China Takes Aim at Comprehensive Mouse Knockout Program

**SHANGHAI**—Geneticist Xiaohui Wu looks through a window into a clean room on the campus of Fudan University here and proudly points to a growing collection of mutant mice. To a visitor, the 4000 cages and 20,000 mice representing 400 mutant strains look pretty impressive. To Wu, the scale of the operation is a frustrating limitation.

"We plan to mutate 70% of the mouse genome over the next 5 years," he says. Yet, their current facilities are filled to capacity. A new building will provide space for 10,000 more cages. But Wu needs 50,000 more, enough for about 100,000 mutant mice. Those cages, he says, require a lot more space and "a lot of money."

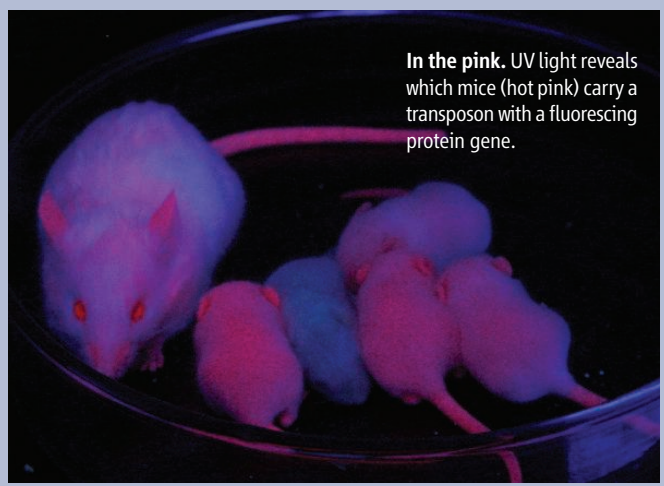
Throughout the world, researchers are setting up programs to shut down the mouse genome gene by gene to learn what each gene does (see main text). The Fudan University mouse facility—a joint effort with Yale University—is shooting to be a key player and hopes to team up with the U.S. National Institutes of Health (NIH) Knockout Mouse Project.

The driving force behind the tentatively named Mammalian Functional Genome Project is Tian Xu, a geneticist at Yale University School of Medicine who is also an adjunct professor at Fudan. The Fudan-Yale group, along with colleagues at the University of Colorado, Boulder, and Duke University in Durham, North Carolina, has come up with an efficient way to knock out mouse genes. They use a transposon, a short segment of DNA that invades genomes, sometimes inserting itself into a gene and deactivating it.

Developmental biologists have used transposons to disable genes in plants, worms, and fruit flies for years, but they had not found one that worked well in mammals. After 8 years of searching, Xu and his colleagues found "piggyBac," which was first identified in the cabbage looper moth by molecular virologist Malcolm Fraser of the University of Notre Dame in Indiana. "We don't know why it works," says Xu. But it does. The group reported its finding in the 12 August 2005 issue of *Cell*.

**Loyal alum.** Yale's Tian Xu and his alma mater are making mutant mice.

The technique is similar to gene trapping in that it randomly disables genes. But using a transposon avoids the laborious manipulation of embryonic stem cells required by other knockout techniques. The researchers made a line of mice that carry both the transposon and DNA that causes the transposon to move. When they mate these mice with wildtype mice, the transposon hops to a new place, preferably to a gene. "All you need to do is just breed mice, and each has different genes mutated," Xu says. This approach can hit genes other knockout approaches tend to miss, he adds.



**In the pink.** UV light reveals which mice (hot pink) carry a transposon with a fluorescing protein gene.

Also, the Fudan-Yale group has put the gene for red fluorescent protein into the transposon. Mice that wind up with the transposon in their genomes are pink under ultraviolet light. "You just look at it, and you can tell" if the genome is carrying the transposon, Xu says.

The Fudan-Yale team opted to set up its large-scale mouse facility in China to save money. Xu estimates that this project could cost one-fifth to one-fourth what it would cost in the United States. But it is still not cheap, and international researchers are impressed by the $12.5 million already pledged from national and local government funding agencies. "I think it's great that [the Chinese] are doing this," says Phil Soriano, a developmental biologist at the Fred Hutchinson Cancer Research Center in Seattle, Washington. Wolfgang Wurst, director of the Institute of Developmental Genetics at the National Research Center for Environment and Health in Munich, Germany, thinks the project is a welcome indication of China's increasingly international orientation. "It is a sign that they are serious research partners," he says.

To leverage support from China itself, Xu and Wu are asking for $30 million from NIH to start mass-producing, preserving, and distributing mutant mice. For the cost of shipping and handling, researchers will receive frozen embryos or sperm, with no intellectual-property–rights restrictions attached. Also, the NIH money would go a long way toward producing the 100,000 strains of transposon-modified mice. Wu and Xu need that number of strains to be sure they have 20,000 genes covered, because the transposon also lands on non-coding regions. If they don't get NIH funding, they may have to recoup some costs by charging fees or placing restrictions on mutant mice, Wu says.

At this point, the other programs are simply making knockout strains, But here, researchers are busy screening the more than 400 mutant mice they have generated over the past year, looking for phenotypes from neurophysiological, immunological, and disease angles, among others. That information will go up on the Web prior to publication, making it easier for potential users to see which mouse will best suit their needs, the duo point out. Four hundred mutants is about the limit until the team's new facility comes on line. After that, the view through these new clean-room windows will get even more interesting. **–DENNIS NORMILE**



needs an animal that wants drugs just because his companions have them.

Setting out to make her dream mouse is not really an option, however, because she has no clue what gene might influence peer-pressure sensitivity. Picciotto might be able to find the mouse in the community after an exhaustive search, but, if it exists, there's a good chance it's tucked away in a cage in a lab somewhere or frozen down as a clump of embryonic stem cells in a biotech company. Either way, it's as good as gone.

Even if Picciotto finds what she is looking for, that's hardly the end of the story. "I'm sorry to say that there are a few labs out there [that] won't share their mice even if they've published them in a journal [such as *Science* or *Nature*] that requires them to do so," says M. Celeste Simon, a developmental and cancer biologist at the University of Pennsylvania Cancer Center in Philadelphia. And as Aguzzi knows all too well, reticent mouse-makers can effectively quash efforts to use their mice by stalling delivery or making outrageous demands about co-authorship.

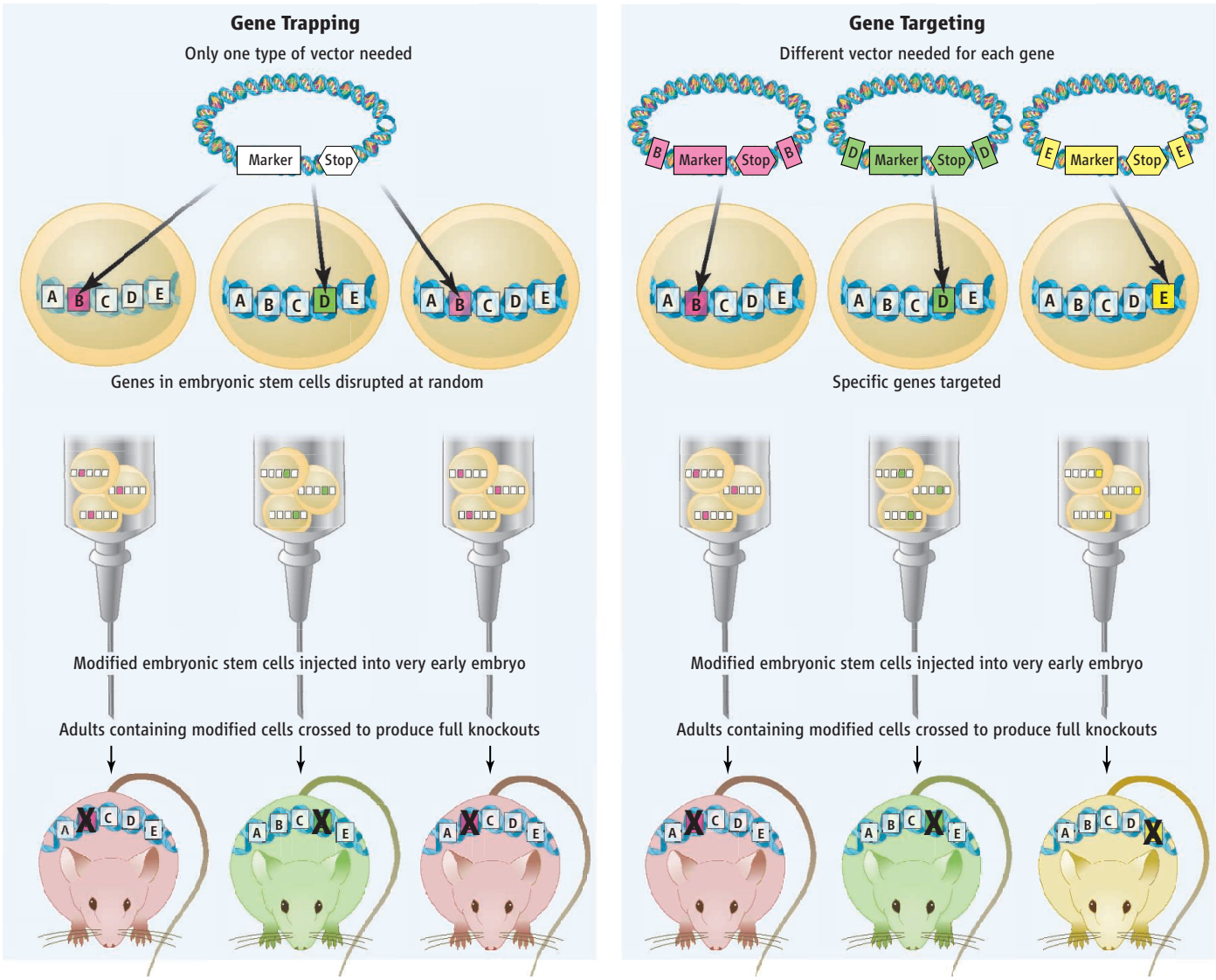



ILLUSTRATION: K. SUTLIFF/*SCIENCE*

**Different strokes.** There's more than one way to knock out a mouse, but each has its pros and cons.

Assuming the source of the mouse is cooperative, "transferring mice is an extremely difficult and time-consuming process," says Simon. Some of Simon's Penn colleagues lost 2 years of work when mice they ordered from a government facility turned out to be infected with an extremely contagious virus that can alter phenotypes. "It strikes fear into one's heart," she says. "Two years is a lifetime in the world of science." Other investigators complain about the cost and hassles of shipping or draconian material transfer agreements.

Over the past 6 years, several efforts have popped up to help address some of these problems. The International Gene Trap Consortium, for example, runs a database that enables researchers to track down about 20% of the existing unique mouse knockouts. And repositories themselves—most of which are publicly funded and store anywhere from 500 to 4000 mice—are beginning to work together under the Federation of International Mouse Resources to help make sure researchers around the world can get any mouse in any repository.

## The big push

Realizing that these were just baby steps, mouse researchers from several countries decided in 2003 to take a giant leap. At a meeting at the Cold Spring Harbor Laboratory in New York, they called for a comprehensive international mouse knockout program. Besides shooting for an IKEA-like superstore, the participants agreed that it would be most economical to avoid trafficking in live mice and instead decided to maintain the knockouts as embryonic stem (ES) cells: clumps of tissue that can be frozen down and later grown up into full-fledged mice. Researchers could request ES cells or be provided with easier-to-use frozen embryos or sperm. They also proposed to use NIH's National Center for Biotechnology Information as their clearinghouse. Its Web site would act as a sort of Google to scan mouse repositories for the desired knockout. "The ultimate goal is to have one-stop shopping [for these mice]," says KOMP Program Director Colin Fletcher.

Two years after the meeting, Wolfgang Wurst, director of the Institute of Developmental Genetics at the German National Research Center for Environment and Health (GSF), and his colleagues set up the European Conditional Mouse Mutagenesis Program (EUCOMM). To get the program rolling, the European Union has promised $16.3 million over the next 3 years. The bulk of the EUCOMM effort is divided between two institutes: GSF and the Sanger Institute. GSF will use "gene trapping" (see diagram, above left) technology to randomly knock out 12,000 genes in ES cells. The Sanger Institute and GSF will use "gene targeting" technology to disable 8000 preselected genes (see diagram, above right).

"It's an ambitious program," says Bradley, who is leading the Sanger effort, "but we're fairly confident we can meet our goals." So far, GSF has produced about 3700 unique knockouts, which researchers can order for $631 apiece. Bradley expects Sanger's lines to start becoming available by late 2007.



Published by AAAS

At the same time EUCOMM was getting started, Canada came out with the North American Conditional Mouse Mutagenesis Project (NorCOMM). Over the next 5 years, Genome Canada will spend $8 million for knockout work primarily at the University of Toronto and the University of Manitoba. The project has produced 3000 gene-trapped knockouts and hopes to make 9000 more over the next 18 months.

NIH's upcoming knockout effort is similar in scope and direction. KOMP expects to spend $50 million at up to four soon-to-be-named centers to build a library of 10,000 knockouts (see sidebar, p. 1863). Like EUCOMM, KOMP will likely use a combination of gene trapping and gene targeting to produce its knockouts. Targeting allows researchers to make precise mutations in their gene of choice, says Fletcher, and targeting will be easier to coordinate among KOMP centers and with the international partners because each group will know exactly what gene it's going after.

But there are important differences between KOMP and the other programs. EUCOMM and NorCOMM are making so-called conditional knockouts, in which the genes that are swapped into the genome have a self-destruct sequence. The new gene encodes information that tells it at which point in development or in which tissue to disappear. The strategy is especially important for determining the function of essential genes, which, if shut off too early, can kill a mouse while it's still an embryo, short-circuiting studies of the gene's effects.

When KOMP knocks out a gene, however, it's dead from day one. More embryos may die than with conditional knockout technology, but these "frank null" knockouts are still very informative, says Fletcher. They tell researchers whether a gene is necessary for development.

Also, of all the mouse efforts, only KOMP will focus on "repatriation." Thanks to NIH's detective work, the agency has compiled a list of the "lost" mice in the community. Recently, in a sort of mouse version of *American Idol*, NIH posted a request asking researchers to vote for the top 20 mice on this list that they'd like to see in a public repository. "That helped us prioritize 500 to 600 mice to repatriate," says Fletcher.

Part of the KOMP effort will involve contacting the owners of these mice and asking them to put their animals in a globally accessible repository. NIH kicked off this program earlier this month, with $800,000 split between the University of California, Davis, and the University of Missouri, Columbia, to acquire 300 of these lines. KOMP leaders hope the repatriation effort will conserve resources by obviating the need to make these lines again.

**Trouble ahead?**

But before a global knockout mouse emporium opens it doors, the international effort must overcome a number of hurdles. Topping the list is figuring out how to avoid the knockout duplication already seen in the mouse community. That's going to be a challenge, especially once each effort is cranking out hundreds of knockouts a month, often in random genes. EUCOMM's Wurst admits it will be "hard to coordinate" his gene-trapping program with NorCOMM's, because neither can predict which genes it's going to knock out. And the American and European groups have yet to factor in the knockouts coming in from China.

Even if redundancy can be addressed, it will still be caveat emptor for researchers who need to compare mice made by different projects. KOMP plans to use a strain of mouse called Black 6, whereas EUCOMM and NorCOMM are making their mutants in strain 129. That could cause studies of behavioral genes, for example, to yield skewed results. "Some 129 strains are really stupid, while Black 6 has a reputation for being smarter," says Yale's Picciotto. "You can't compare the two."

Another unresolved issue is what to do about knockouts that are knockoffs of an already-patented mutant. Several biopharmaceutical companies, including Deltagen in San Carlos, California, make their money selling big-ticket knockout mice. Deltagen, which last year earned $6.7 million from its catalog of 900 knockouts, is seeking "broad patents" on the majority of its lines, says CEO Robert Driscoll. Driscoll would not comment on what steps, if any, the company would take if KOMP or another effort remade one of its patented mice.

On the academic side, some researchers question the way the global endeavor is taking shape. "I'm not totally convinced [this effort] is going about things the right way," says University Hospital of Zurich's Aguzzi. He worries that the variety of strains and technologies being used will lead to glitches in these high-throughput enterprises. The global effort is "layers of magnitude more complicated than the Human Genome Project," he warns.

Aguzzi also emphasizes the need to take one step at a time. He argues that plenty of knockouts have been made with specific biological questions in mind and that these questions should be answered first. "Putting so much effort into creating a bunch of lines that people may not be able to ask the right questions with may not be the best use of resources," he says.

Each effort will try to address this concern by growing a subset of its frozen lines into live mice and then characterizing them. This information will then be uploaded into the central database, so researchers such as Picciotto might find their dream mouse. But a massive phenotyping effort is still years away—the next big step after this big step.

Despite these caveats, the global project should have a dramatic impact on both basic and biomedical research, says Picciotto. "Ordering a mouse is never going to be as easy as ordering an antibody," she says. But as the global project matures and begins to characterize the knockout lines in its libraries, even researchers in small labs and those who are not mouse geneticists will be able to delve into the world of the knockout mouse. "Before, scientists were limited by their experience and their resources," she says. "Now they'll only be limited by their imagination."

**–DAVID GRIMM**


**Gone, but not completely.** Without the *Dicer* gene, a mouse embryo (*inset, left*) is small compared to a normal embryo (*inset, right*) and dies within a week. But when the gene is programmed to turn off just in skin cells, this conditional knockout mouse is born, but has very little hair (*above*).

CREDITS: T. ANDLAND S. E. MILLAR/UNIV. OF PENNSYLVANIA, E. P. MURCHISON AND G. J. HANNON, COLD SPRING HARBOR LABORATORY; (INSET) EMILY BERNSTEIN *ET AL.*, *NATURE GENETICS* **35**, 215–217 (2003)

# MEETINGBRIEFS>>

HUMAN BEHAVIOR AND EVOLUTION SOCIETY | 1–2 JUNE 2006 | PHILADELPHIA, PENNSYLVANIA

## Long-Ago Peoples May Have Been Long in the Tooth

Life may have been nasty and brutish for our prehistoric forebears, but it wasn't necessarily short. Contrary to the notion that it was rare for someone to reach the age of 40 in prehistoric societies, studies of modern hunter-gatherer groups suggest that a substantial proportion of our ancestors survived into their 70s, says an anthropologist who has been studying indigenous people in Bolivia.

**Age old pattern.** A Tsimane senior citizen in Bolivia.

Speaking last week at the meeting of the Human Behavior and Evolution Society in Philadelphia, Pennsylvania, Michael Gurven of the University of California, Santa Barbara, reported on mortality data collected from 10 modern-day groups of hunter-gatherers and forager-horticulturists, including the Tsimane Indians in north-central Bolivia, which he and Hillard Kaplan of the University of New Mexico have studied since 2001.

Gurven's analysis revealed that 40% to 50% of the members of these groups never make it to age 15. But their prospects brighten after that, he says: A 15-year-old has a 40% chance of reaching 65, and by the time they reach 70, the mortality rate is no higher than for a U.S. resident. Evolutionary psychologist Daniel J. Kruger of the University of Michigan, Ann Arbor, says the new work "challenges current thought on … the shape of the hominid survival curve." Gurven and other scientists estimate ages and mortality of contemporary indigenous groups through a variety of convergent techniques, including interviews, old missionary records, historical events, and photographs. In contrast, reconstructions of prehistoric populations rely primarily on skeletal data.

"Some reconstructions of prehistoric populations tend to show life expectancies of 15 to 25 years," Gurven says, "with relatively low infant and child mortality but extremely high adult mortality." Not only is Gurven's work at odds with that scenario, but he says that "adult life expectancy is remarkably similar across these groups."

The research points to an existence structured around longevity. "Adult-level production is controlled by skills and knowledge rather than physical restraints," says Gurven. Although the men in these groups reach the height of their physical powers in their 20s, it is not until their 40s that they reach the peak of hunting prowess, he notes. Rice production by males peaks in their 50s, as they turn from hunting to less rigorous agricultural pursuits. And there is a pattern of resource flow from old to young that continues into the 70s.

Kruger says the research indicates that "once individuals pass the concentrated risk in infancy and early childhood, they are more likely to survive into the older age range (65–75) than previously thought." University of Michigan, Ann Arbor, anthropologist Rachel Caspari, who studies ancient populations, says that it's not possible to know whether prehistoric people were living as long as modern hunter-gatherers without "high-resolution" ways of determining ages of death from fossils. However, she says Gurven's results make sense and that her research supports the notion of a "modern human demographic pattern" that probably has pertained for the past 30,000 years.

Gurven reported that the leading causes of death among those middle-aged and older appear to be from infections and violence. But these indigenous groups share some of the same scourges of aging that affect modern societies. By the age of 65, he says, "almost all have significant pain" from orthopedic problems, especially back pain.

CREDIT: PHOTO COURTESY OF MICHAEL GURVEN

## An Evolutionary Squeeze on Brain Size

Despite the huge individual differences in mental abilities, the size of the human brain varies remarkably little from person to person. In fact, brain volume is evolutionarily more stable than that of any other bodily organ, researchers reported last week. They speculate that our brain, after increasing dramatically in size in early human evolution, ran up against the skull-size limitations imposed by the female birth canal.

Evolutionary psychologists Geoffrey Miller of the University of New Mexico, Albuquerque, and Lars Penke of Humboldt University in Berlin calculated for the human brain a measure called "coefficient of additive genetic variance," or CVA. A formula involving the size range of a physical feature and its heritability, CVA reflects "evolvability": that is, the extent to which the relevant genes are susceptible to change through mutation and natural selection.

The heritability of brain is remarkably high—about 0.9, studies have shown. That makes brain size even more genetically influenced than height, according to Miller. Despite the brain's complexity, he says, "at the genetic level, [brain size] seems as if it's a really simple trait like fruit fly bristle number, … astonishingly ironclad against any environmental perturbation."

Miller notes that he and Penke were surprised at the relatively low CVA for the human brain, because they assumed it had been subject to intense selection. CVAs are likely to be higher for traits that are fitness-related and therefore good candidates for natural selection. But the brain's CVA of 7.8, low for a volumetric trait, means there is limited potential "evolvability." Other features with substantial heritability, such as breasts and kneecaps, have much higher CVAs, said Miller.

Certainly, there has been selection for brain volume in the past, said Miller: As hominids became more intelligent, their brain size tripled over a 2-million-year period to about 1400 cubic centimeters, compared with 370 for chimps. But its growth plateaued about 200,000 years ago, Miller speculates, when it "reached the physical constraint of pelvic size." As a result, he says, brain size "is not a good index of IQ in recent evolution." Adds Penke: "Virtually all theories of brain and intelligence evolution propose a recent history of directional [i.e., 'more is better'] selection for both." But "recent directional selection" on intelligence must have worked on brain features other than absolute size.

That notion is corroborated by brain-imaging studies, says Richard Haier of the University of California, Irvine. Although the correlation between brain size and IQ is only a modest 0.4, he says, the latest imaging studies show much higher correlations of IQ with a "small number of discrete areas" of gray matter.

**–CONSTANCE HOLDEN**



Published by AAAS

# Podosomes and Invadopodia Help Mobile Cells Step Lively

**These feetlike structures aid the necessary migrations of immune and other cells, but also the deadly wanderings of cancer cells**

Anyone who has taken an introductory biology class is familiar with the migratory prowess of the simple amoeba. But many of our own cells could give amoebae a run for their money. Immune cells have to sprint to infection sites to ward off invading pathogens. And cells that help maintain the skeleton nimbly patrol bones. On the minus side, however, cancer cells can travel throughout the body and seed new tumors.

Recently, a couple of hitherto-obscure cellular structures, known as podosomes and invadopodia, have come under increasing scrutiny as possible contributors to such perambulations. These microscopic assemblies, which are possibly related, both form on the bottom surface of the leading edge of migrating cells—a location consistent with the idea that they help cells move. The composition of each reinforces that suspicion: Both contain proteins, such as actin, that have previously been linked to cell motility (*Science*, 10 October 2003, p. 214), as well as enzymes that can break down the proteins of the extracellular matrix (ECM).

Establishing the roles of podosomes and invadopodia hasn't been easy, however, because they have so far been studied in cells maintained on artificial surfaces rather than in live animals. "As far as I know, no one has seen these structures in vivo," notes Gareth Jones of King's College London.

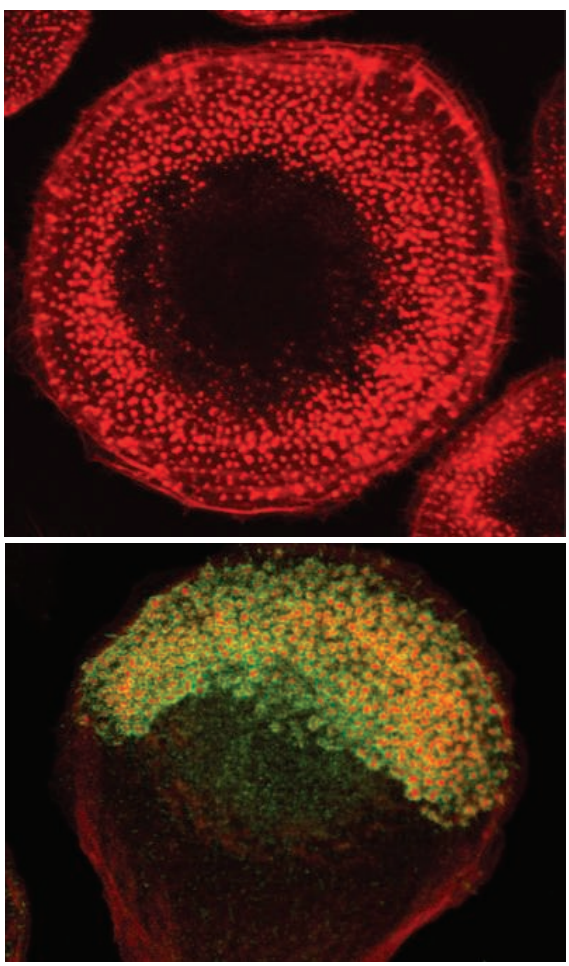
Still, recent research suggests that these cellular features are crucial for understanding and perhaps treating a variety of diseases as diverse as cancer and osteoporosis. Indeed, scientists have already implicated malfunctioning podosomes as a cause of a rare hereditary disease that impairs the immune system. Research on podosomes and invadopodia "is heating up because they've really become physiologically important for things such as metastasis," says cell biologist John Condeelis of Albert Einstein College of Medicine in New York City.

### Feet for cells

Podosomes first appeared on cell biologists' radar screens in the mid-1980s. Pier Carlo Marchisio, currently at San Raffaele Scientific Institute in Milan, Italy, and his colleagues discovered the structures on a variety of migratory cells, including immune cells called macrophages and osteoclasts, which help maintain bone by dissolving away areas that need repair. In all cases, the structures appeared where the cells made contact with a surface. This suggested that they might be involved in cell adhesion, and thus in cell motility, because cells have to stick to the surface over which they are migrating in order to move. Because the structures appeared to act like cellular feet, the Marchisio team coined the name podosomes for them.

Invadopodia came along just a few years later, identified in 1989 by Wen-Tien Chen, now at Stony Brook University in New York. He found that cells transformed by Rous sarcoma virus (RSV) form protrusions at their leading edge when moving on a surface containing ECM proteins such as fibronectin. More intriguing, Chen noticed that holes appeared in the protein substrate at the precise spots where the protrusions were located—an indication that they carry proteases that digest the ECM. "Invadopodia are more than just feet," Chen says. "They have a functional effect." The ability to digest ECM proteins would be very useful to migrating cells, including cancer cells that need to burrow through blood vessel walls in order to spread to distant sites.

Marchisio and his colleagues found similar structures in cells transformed by RSV but considered them to be podosomes. Both researchers now say that they think they were looking at the same thing, yet the precise relationship between podosomes and invadopodia remains a burning issue in the field. Most researchers today define podosomes as more dynamic than invadopodia and smaller—1 to 2 micrometers in diameter compared to 8 to 10 micrometers. But both structures have a core of actin filaments and contain other proteins that regulate actin polymerization. Some researchers, including Roberto Buccione of the Consorzio Mario Negri Sud in Chieti, Italy, and Mark McNiven of the Mayo Clinic in Rochester, Minnesota, have suggested that podosomes form first and then mature into invadopodia, but the jury is still out on this idea.



**Going mobile.** In a moving macrophage (*bottom*), podosomes (stained red for actin and green for vinculin) congregate on the cell's leading edge while they ring the bottom of a quiescent macrophage (*top,* stained red for actin).

### The buzz about WASp

Research on podosomes took off in the late 1990s when they were linked to the then-newly-discovered mutant gene that causes Wiskott-Aldrich syndrome (WAS). Because the gene is on the X chromosome, the disease afflicts mainly boys, causing them to be severely immunodeficient. Without a bone marrow transplant, the children usually die before age 20 of infections or of immune-cell cancers to which they are also prone.

Early analysis of WASp, the protein product of the gene that was found to be mutated in the immune syndrome, suggested that it helps control actin function. Then in the late 1990s, Stefan Linder of Ludwig Maximilians University in Munich, Germany, and his colleagues detected WASp in the actin core of macrophage podosomes. Linder's group and another team led by Jones and Adrian Thrasher of University College London also found evidence that podosome formation requires WASp. "Normal macrophages and dendritic cells have these structures, but cells from boys with Wiskott-Aldrich had none at all. That was a bit of a shock," Jones recalls.

The London team found that the patients' macrophages had also lost the ability to respond to chemokines and cytokines, the chemical signals that normally activate immune cells and attract them to infection sites—a possible explanation for the

CREDITS (TOP TO BOTTOM): IMAGE MADE BY BARBARA SCHROEDER AND STEFAN LINDER; R. BUCCIONE *ET AL.*, *NATURE REVIEWS/MOLECULAR CELL BIOLOGY* 5, 647 (2004), IMAGE BY S. LINDER

impaired immunity of WAS patients. Both podosome formation and the ability to migrate in response to chemical signals were restored by putting the normal *WAS* gene into patient macrophages. "This was the first link [of WASp] to something physiological," Linder says.

Further evidence for a physiological link came from a study that Jones, Thrasher, and their colleagues described in *Blood* in February of last year. Dendritic cells serve as sentinels throughout the body, detecting foreign invaders and initiating immune responses to them. The London team found that the dendritic cells of mice in which the *WAS* gene had been deactivated could not migrate to their normal locations in the lymph nodes, presumably because they can't form podosomes.

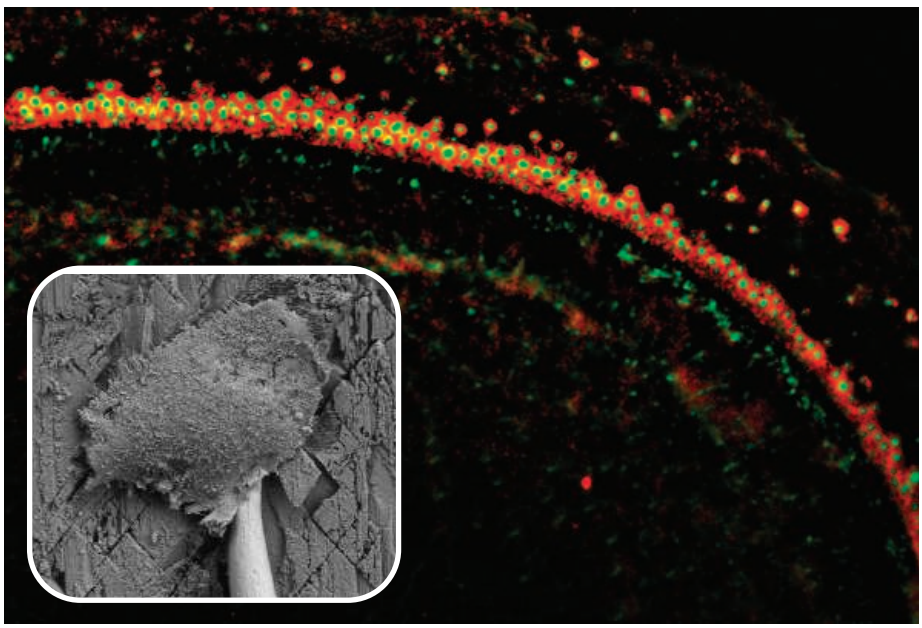
### Cancer connection

The ability to form podosomes and respond to chemokines could have a dark side: Cancer cells may exploit those same skills to spread. In recent work, Condeelis and his colleagues have been teasing out the changes in gene expression that characterize metastatic cancer cells. Having previously shown that the metastatic potential of cancer cells correlates with the cells' ability to migrate toward epidermal growth factor (EGF), the researchers inserted tiny needles containing EGF into mammary tumors in rodents and then analyzed the gene expression patterns of the cells that moved into the needles. These cells had increased expression of numerous genes that promote motility and podosome/invadopodia formation, including WASp. Activation of the genes "determines the willingness and ability [of cancer cells] to move to a portal to escape the tumor," Condeelis says.

Other work by the Albert Einstein team indicates that macrophages and their podosomes are partners in the crime of metastasis. The researchers find that as mammary cancer cells migrate into the blood vessels, they move along with the immune cells. The two cell types pair up because each produces a growth factor that attracts the other: Macrophages associated with the blood vessels attract metastatic cancer cells by secreting EGF, and the cancer cells emit a protein called CSF-1 that attracts more macrophages and stimulates further EGF secretion by the immune cells.

The CSF-1 made by cancer cells also induces in macrophages the expression of WASp and other genes involved in podosome formation; similarly, macrophage-made EGF fosters invadopodia formation by the cancer cells. Macrophages could further help tumor cells penetrate the blood vessel wall; recent work by the Condeelis team has shown that the immune cells' podosomes, like invadopodia, carry enzymes that can digest the ECM. Chen suggests that inhibiting such enzymes, particularly one called seprase that his team has found in the invadopodia of melanoma cells, may thwart metastasis.

In the 1 May issue of *Cancer Research*, Chen's group reports that seprase and a similar enzyme, DPP4, are also located in a complex on invadopodialike structures in endothelial cells. Because tumor growth depends on the creation of new blood vessels by migrating endothelial cells, that finding hints at yet another anticancer strategy: Chen's team blocked endothelial cell migration in lab dishes with monoclonal antibodies targeting the seprase-DPP4 complex.



**Bone repair.** In an osteoclast, podosomes (stained red for paxillin and green for actin) form a belt on the cell periphery. The podosome belt may form the sealing ring that attaches an osteoclast (*inset*) to a bone mineral site that needs repair.

### Boning up on podosomes

The osteoclasts of bone are also migratory cells that may depend on podosomes for their mobility—and perhaps for much more. The cells seek out those areas of bone that need repair. Once there, an osteoclast settles down and forms a structure called a sealing ring, a tight connection between cell and bone. This protects adjacent bone from the osteoclasts. "For an osteoclast to be functional, it needs to isolate its target from the rest of bone; it secretes lots of concentrated acid," says Olivier Destaing of Yale University School of Medicine.

Podosomes appear to be involved in sealing-ring formation. Pierre Jurdic of the École Normale Supérieure de Lyon in France and his colleagues, including Destaing, found that as osteoclasts develop, they first produce individual podosomes that arrange themselves in small rings. These rings then expand to the periphery of the cell, where they form a stable podosome belt. The early view was that the podosome belts might coalesce to form sealing rings, which are thicker, but efforts to detect that transition had produced contradictory results, with some suggesting that the two structures are independent of each other.

Jurdic, Destaing, and their colleagues have found, however, that sealing belts only form when osteoclasts are in contact with apatite, the mineral that forms the solid framework of bone. Apatite somehow activates both the Rho and Src kinase enzymes, both of which are needed for sealing-ring formation. The researchers also found that inhibiting Rho induces the transition of sealing rings into podosome belts.

Further support for the idea that the structures are related comes from Benjamin Geiger's group at the Weizmann Institute of Science in Rehovot, Israel. His team found rearrangements in podosome proteins—in particular, a marked increase in their levels—at locations where sealing-zone structures appeared to be forming from podosome rings. This would be consistent with the notion that sealing rings are thicker versions of podosome belts. If so, drugs that inhibit podosome function in osteoclasts could form the basis of a new osteoporosis therapy; the disorder typically results when the bone-dissolving activities of osteoclasts outpace the bone-forming capacities of their partners, the osteoblasts.

Many questions about podosomes and invadopodia remain to be answered. In particular, researchers are searching for markers that would let them observe the structures in living tissue—and perhaps put to rest the nagging doubts about the relevance of lab-dish studies. That would certainly give them a step up on understanding these cellular feet that are now attracting so much interest.

**–JEAN MARX**





Published by AAAS

# High Court Asks Army Corps To Measure Value of Wetlands

**A divided Supreme Court wants the government to adopt rules on which wetlands deserve federal protection, but scientists say they all matter**

Which wetlands are important enough to protect? That's the question the U.S. Supreme Court put to the U.S. Army Corps of Engineers last week when it ruled on two Michigan cases involving wetlands that property owners wanted to develop. The answer, which will require the corps to define more precisely its jurisdiction under the Clean Water Act, will have ramifications for wetlands across the country.

The 1972 act requires landowners to get a permit for certain actions that might harm wetlands. The corps has claimed an expansive reach that covers any wetland from which water eventually drains into larger lakes and rivers. But on 19 June, in *Rapanos* v. *United States*, the Supreme Court told the corps that's not good enough. For wetlands that aren't next to a navigable river, the key opinion said, the corps should draw up criteria to determine whether they provide significant enough benefits for clean water downstream to be regulated. "It is a critically important decision for our nation's waters," says ecologist Judy Meyer of the University of Georgia, Athens.

The diversity of wetlands makes that task a tough challenge. Meyer and other scientists fear that more-complicated regulations could translate into less wetlands protection, as developers could lobby for some kinds of wetlands to be left out of the rule. In the meantime, lawyers are likely to have a field day as the corps struggles to interpret the Supreme Court's decision on a case-by-case basis.

The lead plaintiff in the Michigan cases, developer John Rapanos of Midland, argued that the corps had no jurisdiction because the wetlands on the contested property were 32 kilometers from navigable waters and connected to them by a mere ditch. Four justices rejected the corps' argument that it can regulate wetlands adjacent to any tributary of navigable water, with Justice Antonin Scalia writing that the corps could only regulate wetlands with a continuous surface connection to "relatively permanent bodies of water." Four other justices took the opposite view, arguing that the corps' existing jurisdiction is reasonable.

That tie set the stage for Justice Anthony Kennedy's decisive vote. He agreed with Scalia that the two cases should go back to the lower courts for further consideration, but he said the corps should have jurisdiction over any wetlands that provide a significant benefit to the "chemical, physical, and biological integrity" of downstream waters. Those whose contributions are "speculative or insubstantial," he wrote, should be outside the corps' purview.

The corps has aimed for rules that are relatively simple and easy to interpret. To decide whether a particular locale falls under its jurisdiction, for example, it has relied upon aerial photographs or topographic maps showing how surface water moves from that wetland to navigable waters. Likewise, the new rule will need to be "easily understandable for the corps analysts and the permittees," says Richard Ambrose, a wetland ecologist at the University of California, Los Angeles. "If you make it too complicated, it will be paralyzing."

**No limits.** Scientists say a new rule should cover even small wetlands, like this Michigan fen, that excel at cleansing water.

CREDIT: STEPHEN HAMILTON

But wetlands—and their functions—resist easy categorization. Their impact on water quality varies by location and other factors; wetlands near an agricultural field, for example, will likely process more nitrate and retain more sediment than wetlands near a pristine forest. And relatively undisturbed wetlands are likely to increase biological diversity in downstream waters, a factor the Clean Water Act is meant to protect. Water levels and flows also vary enormously, ranging from drenched cypress swamps to ephemeral vernal pools. Wetlands in the arid west pose a particular conundrum: When water flows only irregularly, what metrics should be used?

Scientists say rough indicators exist to determine *whether* a wetland is having a positive impact on water quality. Joy Zedler of the University of Wisconsin, Madison, suggests looking at water birds and other aquatic wildlife, the potential to trap sediment and reduce floods, and other factors that can be readily seen or inferred. But those indicators don't answer what Justice Kennedy most wants to know: Which wetlands have enough of an impact on the integrity of waters to qualify for protection? "It is difficult to draw a bright line that works in the practical world of regulation," says wetlands scientist Barbara Bedford of Cornell University.

Even tiny wetlands, like those in headwaters and along small streams, can have a large cumulative impact, studies have shown. A weak current is better than larger streams at trapping silt that would otherwise degrade habitat for salmon and other fish. A slow flow also means that microbes have more time to convert excess fertilizer and prevent downstream algal blooms. In an experimental study published last September in the *Journal of Environmental Quality*, Stefanie Whitmire and Stephen Hamilton of Michigan State University, Hickory Corners, showed that small wetlands in southwestern Michigan were responsible for half of the nitrate removal in the watershed. These benefits diminish when the wetlands are degraded, scientists say.

The corps says it's reviewing the decision and declined further comment. Observers expect the agency to start work right away on interim guidance, although it could take a year or longer to issue a final rule. Until a rule is in place, the courts will proceed case by case. And so will the corps, which each year reviews about 85,000 requests for permits.

Reed Hopper of the Pacific Legal Foundation, which represented Rapanos, has already claimed victory. "The court rejected the idea that there are no limits on the federal government's regulatory authority under the Clean Water Act," he said in a statement. "It is not the role of the federal government to micromanage every pond, puddle, and ditch in our country." But environmentalists say that a seat-of-the-pants approach offers great potential for mischief. "It's an invitation to development interests to contest the corps' authority over wetlands," says Jason Rylander of Defenders of Wildlife in Washington, D.C.

**–ERIK STOKSTAD**

Social science trio
1876

Partnerships to restore degraded lands
1880



# LETTERS

*edited by Etta Kavanagh*

## Testing Climate Reconstructions

A 2005 U.S. CONGRESSIONAL ENQUIRY (*1*) FOCUSED ON THE VALIDITY of the climate reconstruction of the past millennium by Mann *et al.* (*2*) and referred to a *Science* Report that challenged the reconstruction method ("Reconstructing past climate from noisy data," H. von Storch *et al.*, 22 Oct. 2004, p. 679; published online 30 Sept. 2004). This Report was also discussed in the U.S. Senate in 2005 (*3*). In this discussion, it has been overlooked that von Storch *et al.*'s Supporting Online Material (SOM) in fact supports the Mann *et al.* reconstruction.

von Storch *et al.* presented tests of the climate proxy method with two climate models: the HadCM3 model (shown only in the SOM) and the ECHO-G model. Both are compared in the figure. The HadCM3 simulation (solid blue) is consistent with the climate proxy data reconstruction (grey band). The ECHO-G model has since been found to be afflicted by a major artificial climate drift due to an undocumented, inappropriate initialization procedure (*4*).

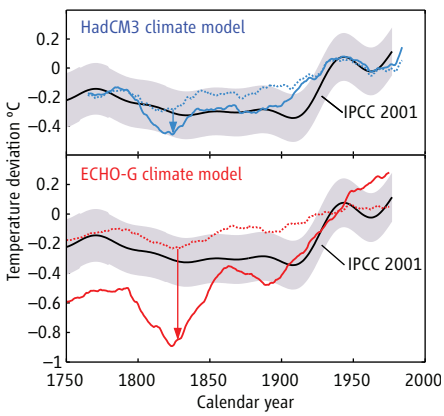
The error of simulated proxies (dotted blue) found in the HadCM3 model is smaller than the error margin given by Mann *et al.* for their method and shown in the IPCC report (*5*). For the time period common to both models, the RMS error of the simulated proxies is 0.24ºC in ECHO-G, but only 0.07ºC in HadCM3—less than one-third.

The two models thus give rather different, conflicting results about the potential errors of proxy reconstructions. This is not mentioned in the Report, which merely states, "Similar results are obtained with a simulation with the third Hadley Centre coupled model (HadCM3), demonstrating that the results obtained here are not dependent on the particular climate characteristics of the ECHO-G simulation" (p. 680).

In addition, it has since been found (*6*) that the proxy method was implemented incorrectly by von Storch *et al.*; with correct implementation, the error is even smaller in HadCM3 than the 0.07ºC shown here. A similar, more recent test with the NCAR climate system model (*7*) also suggests only small errors for the proxy method, supporting the climate reconstruction of the past millennium by Mann *et al*.

**STEFAN RAHMSTORF**

Potsdam Institute for Climate Impact Research, Box 601203, 14412 Potsdam, Germany.




**Test of proxy climate reconstruction method with two climate models, HadCM3 and ECHO-G.** Solid lines show Northern Hemisphere temperature in the models (31-year running means); the dotted lines show simulated proxy reconstructions where the proxies are degraded with 75% noise. The error of the proxy method is the difference between the solid and dotted lines (arrows). For comparison, we show the Mann *et al.* 40-year-smoothed reconstruction for the Northern Hemisphere temperature (black) with its 95% confidence interval (grey), as shown in the IPCC Third Assessment Report (*5*).

### Response

RAHMSTORF CRITICIZES OUR PREVIOUS CONclusions about the climate reconstruction method of Mann *et al.* (*1*) (MBH98). In our previous analyses (*2*, *3*), we found that MBH98 underestimates past temperature variations when tested in climate simulations of the past few centuries. Rahmstorf argues that the simulated Northern Hemisphere temperature lies outside the uncertainty bounds of the pseudoreconstructions in the simulation with the model ECHO-G, but inside the uncertainty bounds in the HadCM3 simulation. He concludes that our analysis supports the Mann *et al.* (*1*) reconstructions. This conclusion is wrong. The problem is the determination of the error bounds.

To successfully compute uncertainty bounds requires an error model. Updated uncertainty bounds for the MBH98 series, on 40-year time scales, can be found in fig. 1B of Gerber *et al.* (*4*). Mann was a co-author on this study, and these uncertainties are consistent with the ones derived in our analysis (*3*). Further, they are about a factor of 3 smaller than those published two years earlier in the IPCC Third Assessment Report (*5*) and used in Rahmstorf's Letter (2σ of roughly 0.07 K rather than 0.25 K for circa 1800). The result of the pseudoreconstruction and the target temperature in the HadCM3 model are therefore statistically well separated when using the proper uncertainties (*3*).

We think that the Letter [as does (*5*)] illustrates a common confusion in our field. There are two sources of uncertainty in reconstructing past climate from proxy records: (i) calibration uncertainty—which part of the signal is not captured by the statistical method; and (ii) residual uncertainty—how much additional, unrelated variability is engraved in the proxy records. Our most recent comment (*3*) did not make this point explicitly, but its uncer-

**References and Notes**

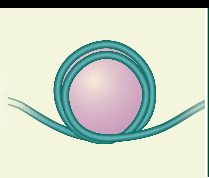
1. See www.realclimate.org/index.php?p=172 for links to the request and the scientists' responses.
2. M. E. Mann, R. S. Bradley, M. K. Hughes, *Geophys. Res. Lett.* **26**, 759 (1999).
3. See http://inhofe.senate.gov/pressreleases/climateupdate.htm.
4. T. J. Osborn, S. T. C. Raper, K. R. Briffa, *Clim. Dyn.* **27**, 185 (2006), DOI: 10.1007/s00382-006-0129-5.
5. IPCC (Intergovernmental Panel on Climate Change), *Climate Change 2001: The Scientific Basis* (Cambridge Univ. Press, Cambridge, 2001), fig. 2.21, p. 134. The error bars for time scales >40 years shown there were computed by Mann *et al.* from calibration residuals, accounting for their spectral "redness." The data were obtained from the National Climate Data Center at http://www.ncdc.noaa.gov/paleo/pubs/mann_99.html.
6. E. R. Wahl, D. M. Ritson, C. M. Amman, *Science* **312**, 529 (2006).
7. M. E. Mann, S. Rutherford, E. Wahl, C. Amman, *J. Clim.* **18**, 4097 (2005).
8. We thank von Storch *et al.* for providing the data of their simulations.



Published by AAAS

Mimicking enzymes

1885



Precision splicing

1886

tainty estimates are based on calibration error. We showed that the MBH98 method implemented in the simulations leads to pseudoreconstructed temperatures being too warm and with differences from the target temperature larger than our calibration uncertainty ranges.

Rahmstorf also alludes to a climate drift in the ECHO-G simulation (*2*). However, this drift mostly affects the earlier centuries of the millennium, when the pseudoreconstruction performs better, and is probably minor after 1400 A.D., when the pseudoreconstruction performs worse. For instance, ECHO-G simulates a difference in the Northern Hemisphere temperature between 1900 and 1980 (the calibration period) and the Late Maunder Minimum (around 1700) of 0.97 K, whereas a simulation with the CSM climate model from NCAR yields 0.87 K (*6*). Therefore, this issue cannot explain the bias of the reconstruction method.

In conclusion, we feel that the paleoreconstruction community would be well served if it used error models describing uncertainties from both calibration and "noise," which leads to uncertainties that have complex, possibly intermittent nonstationary behavior on different time scales. We also urge the community to test methods using realistic "pseudo-proxies" as they offer a good laboratory.

**HAUS VON STORCH,[1] EDUARDO ZORITA,[1] JULIE JONES,[1] FIDEL GONZALEZ-ROUCO,[2] SIMON TETT[3]**

[1]Institute for Coastal Research, GKSS Research Centre, Geesthacht 21502, Germany. [2]Department of Astrophysics and Atmospheric Physics, Universidad Complutense de Madrid, Madrid 28040, Spain. [3]UK Meteorological Office, Hadley Centre, Reading RG6 6BB, UK.

**References**

1. M. E. Mann, R. S. Bradley, M. K. Hughes, *Nature* **392**, 779 (1998).
2. H. von Storch *et al.*, *Science* **306**, 679 (2004).
3. H. von Storch *et al.*, *Science* **312**, 529 (2006).
4. S. Gerber *et al.*, *Clim. Dyn.* **20**, 281 (2003).
5. IPCC (Intergovernmental Panel on Climate Change), *Climate Change 2001: The Scientific Basis* (Cambridge Univ. Press, Cambridge, 2001), fig. 2.21, p. 134.
6. M. E. Mann, S. Rutherford, E. Wahl, C. Ammann, *J. Clim.* **18**, 4097 (2005).

## Team Science and the NIBIB

I WOULD LIKE TO EXPLORE ISSUES RELATING TO the funding of biomedical engineering and imaging at the National Institute of Biomedical Imaging and Bioengineering (NIBIB) that are raised in an article on the Whitaker Foundation ("Spending itself out of existence, Whitaker brings a field to life," D. Grimm, News Focus, 3 Feb., p. 600).

Fiscal year 2005 was the NIBIB's third full year with an operating budget and its third year with a double-digit percentage annual growth in grant applications received. Even in the face of this large growth, the NIBIB's budget projections and management plan resulted in paying to the 20th percentile, well within the range of paylines for the more established institutes. In addition, nearly all of the funded applications contained bioengineering, even those internally labeled as "imaging." Because this type of science is fundamentally interdisciplinary, it is difficult to accurately describe the relative support of biomedical imaging and bioengineering, reflecting progress toward achieving the goal of team science.

In addition to the pervasive bioengineering content in our research portfolio, the institute has multiple training programs that target bioengineering, interdisciplinary science, and young investigators. Indeed, the majority of the NIBIB's current training budget supports such programs. Of note, the NIBIB seeks to significantly enhance the success of new investigators through its policy that increases the payline by 5 percentile points for first-time investigators. During the last fiscal year, this policy resulted in 33% more funded first-time NIBIB investigators.

It is the team science approach, inclusive of young investigators, that is critical for realizing our vision of profoundly improving health care through technological innovation.

**RODERIC I. PETTIGREW**

Director, National Institute of Biomedical Imaging and Bioengineering, National Institutes of Health/U.S. Department of Health and Human Services, Bethesda, MD 20892, USA.

## Funding for Young Investigators at Whitaker

WE WISH TO ADD SOME ADDITIONAL INFORMATION to the points made in the News Focus article "Spending itself out of existence, Whitaker brings a field to life" (D. Grimm, 3 Feb., p. 600). First, it should be noted that nonimaging-related biomedical engineering research was relatively underfunded, not that most of the funding was supporting clinical imaging research. For example, the 2004 numbers indicated that nonimaging research projects constituted less than 40% of the funded individual investigator-initiated grants. Second, the total

research support for biomedical imaging and bioengineering provided by the NIBIB is insufficient to meet the large demands spawned by the Whitaker Foundation in this exploding field. Not only is the NIBIB the second smallest institute, but because of its establishment in 2002, it did not benefit from the doubling of the NIH budget that all other institutes previously enjoyed. Finally, the comment about "not stepping up to the plate" by Nerem referred only to the issue of funding new investigators, which was the specific focus of the young investigator award program of the Whitaker Foundation. For the United States to remain the leader in this highly competitive field, it is important that the NIBIB step in with its own young investigator program to support the bright young people entering the world of biomedical imaging and bioengineering. It is equally important that Congress contribute by providing increased funding for the NIBIB so that the citizens of the United States can reap the benefits of this new area of science and technology.

**ROBERT NEREM[1] AND FRANK YIN[2]**

[1]Institute for Bioengineering and Bioscience, Georgia Institute of Technology, Atlanta, GA 30332–0363, USA. [2]Department of Biomedical Engineering, Washington University, St. Louis, MO 63130, USA.

## Caspase-10 in Mouse or Not?

THE FAMILY OF ASPARTATE-SPECIFIC AND CYSteine-dependent proteases, called caspases, is crucial not only for apoptosis but also for differentiation and cell cycle progression. Several research groups have recently published data ruling in or out the participation of caspase-10, an initiator caspase functioning at the apex of death receptor signaling, in diverse apoptotic processes and in neuroblastoma metastasis (*1*–*5*). Curiously, these conclusions were reached from experiments performed in mouse cells, a species that, according to all public databases and a recent study by Reed and colleagues (*6*), does not contain the *CASP-10* gene in its genome.

How is it possible that these reports still assume the presence of caspase-10 in the mouse? An explanation might be the use of inadequate tools to study processing and activation of caspase-10, such as antibodies and so-called caspase-specific inhibitory peptides. Indeed, all of the above studies used either unspecified caspase-10 antibodies or antibodies that are certified in their respective companies' data sheets to react with cellular extracts of human, mouse, and rat origin. The *CASP-10* gene is also absent from the rat genome. Indeed, a search for commercially available caspase-10 antibodies found that 19 out of 44 caspase-10 antibodies that are distributed by 24 companies are specified on their data sheets to react with mouse or rat tissues. Another seven antibodies were only tested with human cells, and the specificity of at least six other caspase-10 antibodies that supposedly react only with extracts from human cells could be questioned, as they were generated with the same immunogenic peptide as were antibodies that cross-react with mouse and rat tissues.

In addition, many groups based their conclusions about the presence of caspase-10 in the mouse on the utilization of the alleged caspase-10–specific AEVD peptide. However, most caspases including caspase-3 and -8 display an even higher affinity for this substrate ($K_i$ of 42 and 1.6 nM, respectively) than caspase-10 itself ($K_i$, 320 nM) (*7*).

Although the responsibility for this misconception lies clearly in the hands of the individual researchers (and maybe also with the reviewers of the manuscripts), the various companies claiming specificity and applicability of their antibodies in mouse systems are also responsible.

**REINER U. JÄNICKE,* DENNIS SOHN, GUDRUN TOTZKE, KLAUS SCHULZE-OSTHOFF**

Institute of Molecular Medicine, University of Düsseldorf, Universitätsstrasse 1, 40225 Düsseldorf, Germany.

*To whom correspondence should be addressed. E-mail: Janicke@uni-duesseldorf.de

**References**

1. C. Giampietri *et al.*, *Cell Death Differ.* **10**, 175 (2003).
2. R. Kassis *et al.*, *J. Virol.* **78**, 6543 (2004).
3. K. A. Green *et al.*, *J. Biol. Chem.* **279**, 25149 (2004).
4. C. Giampietri *et al.*, *FASEB J.* **20**, 124 (2005).
5. D. G. Stupack *et al.*, *Nature* **439**, 95 (2006).
6. J. C. Reed *et al.*, *Genome Res.* **13**, 1376 (2003).
7. M. Garcia-Calvo *et al.*, *J. Biol. Chem.* **273**, 32608 (1998).

## Role of iNOS in Human Host Defense

IN 2001, S. THOMA-USZYNSKI *ET AL.* WROTE IN *Science*, "in humans the TLR-activated antimicrobial pathway is NO [nitric oxide]–independent" (*1*). In the Report "Toll-like receptor triggering of a vitamin D–mediated human antimicrobial response" (P. T. Liu *et al.*, 24 Mar., p. 1770; published online on 23 Feb.), authors from the same laboratories expanded this view to assert that "antimicrobial activity against intracellular bacteria … in murine, but not human, monocytes and macrophages is mediated principally by nitric oxide" and that this establishes "the evolution of divergent antimicrobial pathways in mice … versus humans. …" The conclusion that humans lack this nitric oxide defense pathway in mononuclear phagocytes is based on in vitro findings that differ in a critical respect from observations of human macrophages in vivo and ex vivo. Thoma-Uszynski *et al.* (*1*) and Liu *et al.* cultured human monocytes under conditions that result in little or no expression of inducible NO synthase (iNOS). In contrast, in vivo and ex vivo, human macrophages do express iNOS in people with infectious and inflammatory diseases (*2*, *3*), notably in tuberculosis.

The role of iNOS in human host defense remains unresolved. The experiments (*4*–*6*) that established the role of iNOS in host defense in mice cannot be performed in people. However, when macrophages expressing iNOS were recovered from patients and infected with mycobacteria in vitro, iNOS inhibitors abolished the macrophages' antimycobacterial activity (*7*).

When a cell type consistently expresses an enzyme in vivo, but differentiates in vitro so that the enzyme is lacking, the cell culture system must be considered deficient as a model. Results in a nonphysiologic cell culture system are not a sound basis for declaring that human evolution has branched off to abandon the use of NO in host defense.

The shortcomings of in vitro systems for human macrophage differentiation are frustrating, particularly because it is so difficult to access human macrophages that have undergone full differentiation and immunologic activation in vivo. Nonetheless, a cell culture model is useful only to the extent that it reflects the biology of the organism. At present, scientists lack the ability to induce iNOS consistently in human macrophages derived in vitro from the monocytes of healthy donors. In vitro studies that truly assess the role of iNOS in human host defense await the development of techniques for iNOS induction in cultured human macrophages that can match that in macrophages of people with disease.

**CARL NATHAN**

Weill Medical College of Cornell University, 1300 York Avenue, New York, NY 10021, USA.

**References**

1. S. Thoma-Uszynski *et al.*, *Science* **291**, 1544 (2001).
2. J. B. Weinberg *et al.*, *Blood* **86**, 1184 (1995).
3. F. C. Fang, *Nat. Rev. Microbiol.* **2**, 820 (2004).
4. J. D. MacMicking *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 5243 (1997).
5. C. A. Scanga *et al.*, *Infect. Immun.* **69**, 7711 (2001).
6. J. L. Flynn, C. A. Scanga, K. E. Tanaka, J. Chan, *J. Immunol.* **160**, 1796 (1998).
7. Y. Nozaki, Y. Hasegawa, S. Ichiyama, I. Nakashima, K. Shimokata, *Infect. Immun.* **65**, 3644 (1997).

### Response

A KEY POINT IN OUR RECENT PAPER WAS ONLY partially quoted by Nathan. The complete quote follows: "In innate immune responses, activation of Toll-like receptors (TLRs) triggers direct antimicrobial activity against intracellular bacteria, which in murine, but not human, monocytes and macrophages is mediated principally by nitric oxide." Our data do establish mechanisms by which activation of the innate immune system via TLRs leads to antimicrobial activity, and in this context, it is not unreasonable to suggest "the evolution of divergent antimicrobial pathways in mice (nocturnal animals that use nitric oxide) versus

humans (daytime creatures that synthesize vitamin $D_3$ in the skin on exposure to UV light)." We were careful to state in the paper that, "We do not imply that this is the only antimicrobial mechanism available to human macrophages." Nathan's research on nitric oxide in acquired immunity is clearly important and is cited in our paper. We look forward to learning the mechanism by which nitric oxide is activated in human macrophages and contributes to immunity to tuberculosis.

**PHILIP LIU,[1] JOHN S. ADAMS,[2] ROBERT L. MODLIN[1]**

[1]Department of Microbiology, Immunology, and Molecular Genetics, and Division of Dermatology, Department of Medicine, David Geffen School of Medicine, University of California at Los Angeles, Los Angeles, CA 90095, USA. [2]Department of Medicine, Division of Endocrinology, Cedars-Sinai Medical Center, Los Angeles, CA 90048, USA.

## TECHNICAL COMMENT ABSTRACTS

### COMMENT ON "Ancient DNA from the First European Farmers in 7500-Year-Old Neolithic Sites"

**Albert J. Ammerman, Ron Pinhasi, Eszter Bánffy**

On the basis of analysis of ancient DNA from early European farmers, Haak *et al*. (Reports, 11 November 2005, p. 1016) argued for the Paleolithic ancestry of modern Europeans. We stress that the study is more limited in scope than the authors claim, in part because not all of the skeletal samples date to the time of the Neolithic transition in a given area of Europe.

Full text at www.sciencemag.org/cgi/content/full/312/5782/1875a

### RESPONSE TO COMMENT ON "Ancient DNA from the First European Farmers in 7500-Year-Old Neolithic Sites"

**Joachim Burger, Detlef Gronenborn, Peter Forster, Shuichi Matsumura, Barbara Bramanti, Wolfgang Haak**

The discovery of mitochondrial type N1a in Central European Neolithic skeletons at a high frequency enabled us to answer the question of whether the modern population is maternally descended from the early farmers, instead of addressing the traditional question of the origin of early European farmers.

Full text at www.sciencemag.org/cgi/content/full/312/5782/1875b

## Letters to the Editor

Letters (~300 words) discuss material published in *Science* in the previous 6 months or issues of general interest. They can be submitted through the Web (www.submit2science.org) or by regular mail (1200 New York Ave., NW, Washington, DC 20005, USA). Letters are not acknowledged upon receipt, nor are authors generally consulted before publication. Whether published in full or in part, letters are subject to editing for clarity and space.

POLITICAL PSYCHOLOGY

# The Perils of Prognostication

**John T. Jost**

We are all dilettantes in many, if not most, areas of life and learning. When we ponder possible futures and appropriate courses of action and we encounter the limits of our own understanding, what can we do but turn to the experts on matters ranging from the weather and the stock market to the health of our bodies and our nations and so much in between? We realize (at least sometimes) that we don't know what the future holds, but at least the experts have a pretty good idea. Don't they?

For anyone who gains solace or inspiration from the conviction displayed by Sunday morning political pundits or "I told you so" Monday morning quarterbacks that populate every field, Philip Tetlock's *Expert Political Judgment* will be sobering. The results of his painstaking research are complex, nuanced, and contingent, but the bottom line is clear enough. Tetlock's data "plunk human forecasters into an unflattering spot along the performance continuum, distressingly closer to the chimp than to the formal statistical models." In fact, "*it is impossible to find any domain* in which humans clearly outperformed crude extrapolation algorithms, less still sophisticated statistical ones" (emphasis in original). Worst of all, those experts with the poorest track records are the most likely to show up on TV screens and blogsites everywhere.

Tetlock is a social psychologist by training, a political scientist by choice, and now a business school professor (at the University of California, Berkeley) by avocation. For over 20 years, he has been a pioneer in the relatively young interdisciplinary field of political psychology. His wide-ranging, partially overlapping interests in lay theories of epistemology and philosophy of science, cognitive styles, motivated reasoning, political ideology, domestic and foreign policy decision-making, counterfactual thinking, and accountability are all brought together in this, his most ambitious, profound, and integrative book to date. In many ways, it is a tour de force, providing as it does a vivid, sophisticated illustration of our limitations in forecasting and, at the same time, the analytical power of our psychological tools when applied in retrospect.

Tetlock asked 284 experts with advanced educational and professional training in international relations, political science, law, economics, business, public policy, and journalism to make thousands of predictions between 1988 and 2003. Participants rendered both short-term and long-term subjective probability estimates of hypothetical events that were inside and outside their domain of expertise, including the Persian Gulf War, the transition from Communism in Eastern bloc countries, the fall of apartheid in South Africa, the outcomes of U.S. presidential elections, the existence of weapons of mass destruction, and the bursting of the Internet bubble. These topics are so intriguing that one wants to see detailed information concerning their predictions on a case-by-case basis. Unfortunately, Tetlock keeps the reader fairly removed from the raw, unprocessed data and offers instead more abstract generalizations concerning the characteristics of better and worse judges.

**Expert Political Judgment**
How Good Is It? How Can We Know?

***by Philip E. Tetlock***

Princeton University Press, Princeton, NJ, 2006. 337 pp. $35, £22.95. ISBN 0-691-12302-2. Paper, $19.95, £12.95. ISBN 0-691-12871-5.


To cope with the mind-boggling complexity involved in processing over 80,000 expert predictions and distilling the concomitants of accuracy, Tetlock boils things down to a single dimension of cognitive style that captures most of the good judgment he could find. Drawing on an essay by Isaiah Berlin, Tetlock distinguishes between "foxes," who "'know many little things,' draw from an eclectic array of traditions, and accept ambiguity and contradiction as inevitable features of life" and "hedgehogs," who "'know one big thing,' toil devotedly within one tradition, and reach for formulaic solutions to ill-defined problems."

How does Tetlock measure the location of each of his experts on the fox-hedgehog continuum? In the book's Methodological Appendix, we learn that he used a factor analysis of responses to a "styles of reasoning" questionnaire comprising 13 items. Eight items were drawn from the "need for cognitive closure" scale (*1*). One item provides respondents with Berlin's definition and asks them to classify themselves as either foxes or hedgehogs. The remaining four items focus on relative preferences for simplicity, parsimony, predictability, and decisiveness (all of which are more appealing to hedgehogs than to foxes). Tetlock's hedgehog-fox score is based on the seven items that had the highest loadings (above 0.25) on the first factor.

Much of the book details the ways in which foxes outperform hedgehogs as prognosticators and Bayesian updaters. Foxes scored higher than others on measures of calibration; their subjective probability estimates were better correlated with the objective frequencies of the events they were predicting, especially in the short term. The worst judges were hedgehog extremists who made long-term predictions in their own areas of expertise. They correctly anticipated war in the former Yugoslavia, but they also predicted several wars that did not happen. Even more than others, they frequently overestimated the likelihood of drastic changes from the status quo.

When unexpected outcomes occurred, hedgehogs were less likely than foxes to revise their beliefs in light of new realities. They were also more likely to display hindsight bias, believing that they "knew it all along," even when they did not, and they were less charitable toward their competition, exaggerating the extent to which rivals were mistaken. The only advantage hedgehogs enjoyed—other than greater media exposure—was a tendency to swing for the home-run fences. They were almost twice as likely as foxes to declare certain events as either inevitable or impossible, and when they did so they were usually correct.

There is an emerging (or perhaps reemerging) scientific controversy concerning the nature of the relationship between political ideology and general psychological character-

CREDIT: JOE SUTLIFF

The reviewer is at the Department of Psychology, New York University, 6 Washington Place, 5th Floor, New York, NY 10003, USA. E-mail: john.jost@nyu.edu

istics. Much evidence—including results from Tetlock's own research (*2, 3*)—indicates that personal needs for order, structure, and closure are positively associated with conservatism, whereas openness, tolerance for ambiguity, and integrative complexity are positively associated with liberalism (*4*). Others deny that these associations are important or consequential (*5*). Given the strong overlap in how the fox-hedgehog dimension and these other variables are measured, there is a missed opportunity to investigate in detail the effects of ideology (as mediated by cognitive style) on judgment and prediction. Tetlock suggests that his fox-hedgehog dimension is unrelated to the left-right ideological dimension, but he does not provide direct or sufficient information bearing on the nature of their association.

Nevertheless, as Tetlock points out, the ideological range of his expert sample may not have been wide enough to adequately test the "rigidity-of-the-right" hypothesis. And in any case, it does seem likely that leftists would be better predictors in some domains and rightists in others. Given that people see the future (at least in part) as they would like it to be, an answer to the question of whether liberals or conservatives are more accurate in their predictions depends upon, among other things, whether the world happens to turn to the left or right during the specified time period.

Tetlock does illustrate, quite cleverly, the effects of ideology on perceptions of historical counterfactuals. Whereas conservatives were convinced that the Soviet Union would not have changed without external pressure from the West (and liberals were more optimistic about internally generated reform), liberals believed that apartheid would not have ended without Western sanctions (and conservatives found it more likely that change could have come purely from within South Africa). Taking a staple from classic social psychology, Tetlock concludes that, "The operative principle is dissonance reduction: the more we hate a regime, the more repugnant it becomes to attribute anything good to redemptive dispositions of the regime (such as a capacity for self-correction)."

The strength of the analysis presented in *Expert Political Judgment* lies in the author's carving out the copious space between normative standards of prediction captured by complex equations requiring difficult-to-calculate base rates and statistical modeling of stochastic processes, on one hand, and descriptive evidence concerning the actual predictions made not merely by ordinary human beings but by the most talented experts among us, on the other. To fill in that space, Tetlock adroitly wields a succession of theories and findings from social and cognitive psychology. In the process he advances considerably the important work begun by Daniel Kahneman and Amos Tversky (*6*), demonstrating just how desperately we need a scientific psychology of judgment and decision-making to correct for the many failings of a ruminating species.

**References**

1. See, e.g., D. M. Webster, A. W. Kruglanski, *J. Pers. Soc. Psychol.* **67**, 1049 (1994).
2. P. E. Tetlock, *J. Pers. Soc. Psychol.* **45**, 118 (1983).
3. P. E. Tetlock, *J. Pers. Soc. Psychol.* **46**, 365 (1984).
4. J. T. Jost, J. Glaser, A.W. Kruglanski, F. Sulloway, *Psychol. Bull.* **129**, 339 (2003).
5. J. Greenberg, E. Jonas, *Psychol. Bull.* **129**, 376 (2003).
6. D. Kahneman, A. Tversky, *Econometrica* **47**, 263 (1979).

10.1126/science.1129403

SOCIOLOGY

# Fishing Rights and Race Relations

**Nicholas J. G. Winter**

Scholarship on whites' opinions on matters of race in America has been fueled by an apparent contradiction. On the one hand, there has been a dramatic long-term trend among whites away from support for overt racism. On the other hand, support for policies intended to address racial inequities—such as school busing or affirmative action—has not increased. There is a disjunction between genuine support for the abstract principle of racial equality and ambivalence or opposition to policies that seem logically connected to that principle.

The literature provides three basic explanations for this pattern. The first suggests that many whites oppose these policies not for racial motives but because they see such interventions as undermining cherished American values such as individualism (*1*, *2*). If so, the contradiction between support for racial equality and opposition to policies is only apparent, because that opposition is not ultimately racial. The other two approaches draw attention to the role of race, but in rather different ways. The realistic group conflict school suggests that white opposition to policies is based importantly in material racial group interest: when policies hurt whites as a group, whites oppose them (*3*, *4*). The symbolic racism school argues that although whites no longer generally endorse traditional anti-black stereotypes, a new form of racism has evolved that condemns blacks and other subordinate racial groups for their perceived failure to live up to American values like individualism and the work ethic (*5*, *6*). This new symbolic racism represents a blending of anti-black feelings with these values. There is long-running debate among proponents of these approaches—and important variants of all three—that has turned on questions of theory, measurement, and statistical methodology.

*Prejudice in Politics* moves beyond the traditional focus on black-white relations to explore these questions in the context of the controversy over Chippewa fishing and hunting rights in northern Wisconsin. Litigation over the treaties that ensure these rights ran from the mid-1970s through the early 1990s and spawned an intense and extended period of racial conflict. Lawrence Bobo and Mia Tuan (sociologists at Stanford University and the University of Oregon, respectively) offer an important contribution to the racial attitudes literature. The book is also an excellent resource for a broader audience interested in the continuing role of race and racism in American society and politics.

Drawing on sociologist Herbert Blumer's classic work on race relations (*7*), Bobo and Tuan develop a "group position" model of racial attitudes that integrates aspects of both the group conflict and the symbolic racism approaches. They take seriously the role of individual-level prejudice, while also arguing convincingly that the social, historical, and political processes that create and institutionalize racial group differences are important in determining racial attitudes. They argue that racial groups are more than "mere" groups: they are an important way that society allocates rights to "scarce and socially valued goods and resources." The construction and maintenance of racial categories are important ways that appropriate roles, rewards, and outcomes are delineated in society. When that system is threatened symbolically—as in the treaty rights dispute—reactions by the dominant group can be strong and well out of proportion to the actual material threat, which was, in fact, minimal in this case. Bobo and Tuan's compelling theoretical development is pleasingly accessible to those not versed in the literature on racial attitudes. At the same time the authors provide signals (and references) that allow specialists to place their approach in the context of the broader literature.

**Prejudice in Politics**
Group Position, Public Opinion, and the Wisconsin Treaty Rights Dispute

***by Lawrence D. Bobo and Mia Tuan***

Harvard University Press, Cambridge, MA, 2006. 288 pp. $40, £25.95, €36.90. ISBN 0-674-01329-8.

The empirical heart of the book is based on a 1990 public opinion survey of Wisconsin residents, which allows Bobo and Tuan to document white opinion (*8*) on this serious racial

The reviewer is at the Department of Politics, University of Virginia, Charlottesville, VA 22904, USA. E-mail: nwinter@virginia.edu

**Prejudiced response.** Efforts by the Chippewa to exercise their treaty rights to spearfishing in northern Wisconsin met with racially charged protests from whites.

dispute. The data also allow them to develop measures of the key constructs from each of the three contending racial attitudes models. Their results compellingly demonstrate the failure of the nonracial values approach to explain white opinion. The authors present a series of statistical analyses that demonstrate the impact of racial predispositions on opinion, above and beyond individual demographic characteristics. They also make excellent and extensive use of respondents' own words, from open-ended responses, to show the ways that white Wisconsin residents' reactions to the Chippewa and to treaty rights are deeply and subtly inflected with racial considerations. These findings underline the conclusion that matters of race are still very much a part of white Americans' political cognition.

The data are less able to distinguish between the group position and the symbolic racism models. Both models deal with racial prejudice, with important—though subtle—differences in their understanding of prejudice. Bobo and Tuan present compelling evidence that the empirical data are consistent with their group position model. As they acknowledge, however, the data are not incompatible with the symbolic politics model. I believe this is not a failure in their choice of this case study or in the development of their survey questions. Rather, the theoretical distinctions being drawn in the modern versions of these various explanations are fine enough that survey data are hard-pressed to distinguish among them. The measures of the building blocks of the models—stereotyping, group competition, political or group threat, symbolic racism, group affect—are too highly correlated to allow a convincing winner to emerge from head-to-head statistical competition. This means that the authors' ability to adjudicate between group position and symbolic racism models turns importantly on a subtle reading of the open-ended data.

Despite this, *Prejudice in Politics* (along with work in the symbolic racism tradition) has important lessons for our understanding of American democracy broadly speaking. There is a long tradition, dating back at least to Alexis de Tocqueville and Gunnar Myrdal, of seeing white Americans' opposition to the advancement of racial "others" as mere irrational prejudice, fundamentally unconnected to the true essence of American culture, society, and politics. Bobo and Tuan show that whites' attitudes are to a considerable extent based on racial predispositions and that those predispositions represent far more than irrational individual dislike. Rather, Americans' racial attitudes connect importantly with the ways that racial categories are constructed and institutionalized in social structure and political conflict. In this sense, they are a fundamental—if distasteful—part of American society and culture.


**References and Notes**

1. P. M. Sniderman, T. L. Piazza, *The Scar of Race* (Harvard Univ. Press, Cambridge, MA, 1993).
2. P. M. Sniderman, E. G. Carmines, *Reaching Beyond Race* (Harvard Univ. Press, Cambridge, MA, 1997).
3. J. M. Glaser, *J. Polit.* **56**, 21 (1994).
4. L. Bobo, in *Eliminating Racism: Profiles in Controversy*, P. A. Katz, Ed. (Plenum, New York, 1988), pp. 85–116.
5. D. O. Sears, in *Eliminating Racism: Profiles in Controversy*, P. A. Katz, Ed. (Plenum, New York, 1988), pp. 53–84.
6. D. R. Kinder, L. M. Sanders, *Divided by Color: Racial Politics and Democratic Ideals* (Univ. Chicago Press, Chicago, 1996).
7. H. Blumer, *Pac. Sociol. Rev.* **1**, 3 (1958).
8. The authors report that they "were unable to sample opinions among any significant number of American Indians, including the Chippewa themselves."


10.1126/science.1129404

POLITICAL SCIENCE

# Stoking the Voters' Passions

**James N. Druckman**

In its defense of the United States Constitution, *The Federalist Papers* make clear that input from citizens must be limited because they think too emotionally. *Federalist* 49 states, "The danger of disturbing the public tranquility by interesting too strongly the public passions, is a still more serious objection against a frequent reference of constitutional questions to the decision of the whole society." Further on, toward the end of the essay, the author (Alexander Hamilton or James Madison) concludes, "The *passions*, therefore, not the *reason*, of the public would sit in judgment. But it is the reason, alone, of the public, that ought to control and regulate the government. The passions ought to be controlled and regulated by the government" (*1*).

**Campaigning for Hearts and Minds**
How Emotional Appeals in Political Ads Work
***by Ted Brader***
University of Chicago Press, Chicago, 2006. 296 pp. $60, £38. ISBN 0-226-06988-5. Paper, $24, £15.50. ISBN 0-226-06989-3. Studies in Communication, Media, and Public Opinion.

Over two centuries later, this view continues to be the conventional wisdom for many. Social scientists, however, have offered little insight into the role of emotion in shaping citizens' political decisions. Do emotions play a substantial role? If so, when? And is such a role problematic? With each technological innovation in the mass media that offers politicians new means to play on the public's emotions, these questions become more pressing. It is such questions that frame the topics Ted Brader addresses in *Campaigning for Hearts and Minds*.

Brader, an assistant professor at the University of Michigan, begins by noting the development of two recent but largely distinct research programs in political science. One focuses on how mass communication affects citizens' opinions. Using content analyses, experiments, surveys, and case studies, social scientists from various disciplines have shown—not surprisingly—that what politicians and news sources say can shape what citizens think and believe. Another fairly recent body of work shows citizens' actions and opin-

The reviewer is at the Department of Political Science, Northwestern University, Scott Hall, 601 University Place, Evanston, IL 60208, USA. E-mail: druckman@northwestern.edu

CREDIT: C. HOYT/MILWAUKEE JOURNAL SENTINEL

ions depend in fundamental ways on their emotions. For example George E. Marcus, W. Russell Neuman, and Michael MacKuen have shown that people tend to become more politically involved and interested when they feel enthusiastic and tend to become more attentive, information-seeking, and open to attitude change when they feel anxious (*2*).

Brader impressively brings these two programs together to probe the age-old concern of whether politicians can manipulate emotional whims to their advantage. He specifically focuses on television ads, which, he explains, "have become the principal tool of contemporary electioneering" in the United States. After reviewing prior related work and assessing "conventional wisdoms" about advertisements and emotions, Brader offers a psychological theory of emotional appeals.


**Cueing emotion.** In Brader's enthusiasm experiment, one ad contained hopeful images and music while the other relied on the narration and less evocative imagery. (Both ads used the same positive script, here: "There's good news in your neighborhood. The future looks bright for a generation of young people.") The fear experiment compared the effects of the same unevocative imagery with those of threatening images and dissonant music. (In the negative narration for this pair, "It's happening right now in your neighborhood. A generation of young people is in danger.")

CREDIT: COURTESY TED BRADER

Brader extends Marcus and colleagues' aforementioned theory of affective intelligence (*2*) and research in psychology by Richard Lazarus, Jeffrey Gray, David Watson, and others (*3*–*5*) by incorporating the effect of political communication (e.g., advertisements) on different emotions and, consequentially, attitudes and behaviors. For example, ads that generate enthusiasm will increase political interest, participation, and confidence, whereas fear-provoking ads will cause people to reevaluate their preferences and potentially change their opinions.

Perhaps the book's major contribution is to then describe what types of advertisements stimulate enthusiasm or fear, test the impact of such advertisements, and demonstrate the relevance of these advertisements to ongoing political campaigns. The author presents results from a set of compelling experiments that he implemented during the 1998 Massachusetts gubernatorial primary. Brader recruited a diverse set of participants to watch a prerecorded segment of nightly news along with the accompanying commercials (in a comfortable setting that resembled a living room). He then randomly inserted different versions of a single candidate advertisement into the commercials. Thus participants randomly saw either no advertisement, one of two ads that had little emotional content, or one of those same two ads but with elements meant to stimulate enthusiasm or fear. (He randomly varied other aspects of the ads such as the sponsor of the ad.) Importantly, the elements he used to stimulate enthusiasm or fear have nothing to do with the ads' contents; rather, he manipulated emotional stimulation entirely by including or excluding certain images and music. For example, the enthusiasm ad added uplifting music and brightly colored images of children playing and smiling, whereas the no-enthusiasm ad had no music and used distant, expressionless pictures.

Although there are some unexpected results, the bottom line is that overall Brader's evidence strongly supports his expectations. Enthusiastic ads motivate individuals to participate (e.g., willingness to volunteer, intention to vote), and once participating, these individuals are likely to become even more committed to their prior preferences. The implication is that enthusiasm leads to political polarization by pushing voters to take action on behalf of their prior convictions. Fear ads have less participatory power—although to some extent they motivate sophisticated individuals. But, fear can open the gates of persuasion, and these ads tend to cause individuals to consider new information and possibly change their political preferences. These findings have important normative implications. Even though citizens clearly act on emotions when prompted to do so, their actions need not result in "bad" outcomes. Indeed, possible consequences of the use of emotional ads include a more active (enthusiastic) populace and citizens who seek new information (to temper their fear).

Brader follows up his experimental findings with a comprehensive content analysis of over 1400 candidate ads from 1999–2000. He shows that in more than three-quarters of their ads, candidates aim to stimulate enthusiasm, fear, or some other emotion through the use of music, color, and other visual cues. The author also demonstrates that emotional ads do not generally lack logic, facts, or policy discussion. Emotion should not be equated with a lack of substance.

Brader's experiments, content analysis, and data presentations are careful and methodical. He explicitly addresses a number of complications, and *Campaigning for Hearts and Minds* shows the methodological state of the art in political communication research. The book also reflects the current status of research on emotion in political science. And this raises an important question: Should political scientists delve beyond behavioral manifestations of emotion to explore processing? Advances in neuroscience make it possible to focus on much more precise processes. If social scientists opt to take this route, which certainly offers some benefits, it is crucial they do not neglect the political, economic, and social environments that define their disciplines. Indeed, another feature missing from most research into political communication is attention to common elements of the political contexts, such as competition among messages (e.g., advertisements) over time. One challenge for the next generation of political communication research is to incorporate such political contexts while probing deeper into the underlying psychological mechanisms.

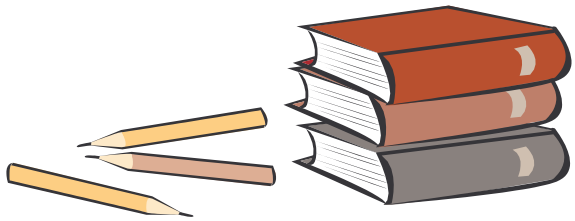
**References and Notes**

1. An online copy of *Federalist* 49 is at http://thomas.loc.gov/home/histdox/fed_49.html.
2. G. E. Marcus, W. R. Neuman, M. MacKuen, *Affective Intelligence and Political Judgment* (Univ. Chicago Press, Chicago, 2000).
3. R. S. Lazarus, *Emotion and Adaptation* (Oxford Univ. Press, Oxford, 1991).
4. D. Watson, L. A. Clark, A. Tellegen, *J. Pers. Soc. Psychol.* **54**, 1063 (1988).
5. J. A. Gray, *The Psychology of Fear and Stress* (Cambridge Univ. Press, Cambridge, ed. 2, 1987).

10.1126/science.1129514

COMMUNITY COLLABORATIONS

# Collaborative Ecological Restoration



Warren Gold,[1]* Kern Ewing,[2] John Banks,[4] Martha Groom,[1] Tom Hinckley,[2] David Secord,[3] Daniela Shebitz[2]

Faculty and students at the University of Washington forge interdisciplinary partnerships with the regional community to restore damaged habitats.

The complexity of the interface between human communities and ecological sustainability demands that we supersede our traditional, balkanized disciplines (*1*). The field of ecological restoration showcases the necessity and merits of interdisciplinary approaches to real-world problems. Drawing on ecology, other life sciences, physical and social sciences, and the humanities, ecological restoration has a long history (*2*).

The dust bowl of the midwestern United States in the 1930s was rehabilitated with the benefit of diverse skills that were grounded in ecological science. Subsequent experience in treating degraded landscapes has developed the field of ecological restoration with a focus on returning biological potential and ecological integrity to damaged land (*3*, *4*) [supporting online material (SOM) text]. Successful restoration requires interdisciplinary participation from land managers, policy-makers, scientists, and educators (*5*, *6*).

In 1999, the University of Washington (UW) began linking components of restoration ecology studies across academic units among its three campuses (*7*). The UW Restoration Ecology Network (UW-REN) was created with one-time internal funds to catalyze faculty and student participation across the traditional boundaries of academic departments (*8*).

Students from natural and social sciences and humanities can earn an academic certificate (similar to a minor) in Restoration Ecology on any of the three UW campuses. Students learn how knowledge from their discipline applies to restoration. All students participate in a year-long ecological restoration project. This capstone project engages students in interactive hands-on learning, revealing the complexity of real-world solutions and creating bonds between the university and the public (see photograph, below).

**Retrieving degraded land.** UW-REN capstone students restoring streamside habitat.

## UW-REN Capstone Projects

The restoration ecology capstone lasts for one academic year (fig. S1) and is directed by diverse faculty (*9*). Students, ideally representing a range of scientific and humanities fields, team up for aquatic and terrestrial restoration projects in neighboring communities. Participation is limited to senior students, who have sufficient knowledge of their field to contribute to a multidisciplinary team.

Projects are selected by a panel of UW-REN faculty that meets prospective clients, who might represent local governments, schools, utilities, foundations, or community groups (table S1). Usually, projects are proposed by clients who would otherwise be working independently or with private-sector consultants but lack the financial or technical resources to do so. Some clients do have the resources but prefer to work with our students. Projects are selected on the basis of their regional ecological importance and potential for community and client involvement. These projects are also chosen for their educational value in representing a range of ecological and restoration challenges, and they must be feasible in size and scope for a student team to handle in an academic year. Project sites have generally been less than one acre, although larger project sites have been restored by multiple student teams working in sequential years.

## The Capstone Experience

The first course, in the autumn quarter, introduces restoration tools through lectures, field visits, and demonstrations by restoration professionals. Topics range from scientific (site assessment, bioengineering, invasive species, and ecological theory application) to pragmatic (project management, grants, and community relations). Previous projects are studied to judge how well original goals were achieved and to understand the sustainability of outcomes. Teams of four to six students make their initial visit to the project site. This quarter builds shared foundations of scholarly and practical knowledge and promotes interdisciplinary communication in the teams.

The second course, in the winter quarter, begins with an ecological site analysis during which students collect the information necessary to design and implement their project. Student teams then draft a proposal for the client, responding to the client's stated needs by describing proposed actions and the ecological possibilities of the site. The proposal undergoes peer review by students in other capstone teams and UW-REN faculty before it is sent to the client. Next, by discussing and clarifying objectives and budget constraints with the client, the team negotiates a final proposal. A detailed work plan is then created that is rigorously tied to client needs, site conditions, prospects for long-term stewardship, and the underlying science. Students draw on their academic backgrounds to connect site-specific project implementation to previous knowledge represented in the literature (see table). That each restoration site is unique complicates this task but also proves the utility of providing a process to formulate unique responses.

In the spring quarter, teams finish site preparation, control invasive species, install habitat elements (including plants), describe the baseline conditions after installation, and develop long-term maintenance strategies. Teams craft stewardship plans, train their clients in the plan, and galvanize community support to ensure long-term project success. All project documents are accessible online. A poster session, attended by current and former clients, students,

CREDIT: WARREN GOLD

[1]Interdisciplinary Arts and Sciences Program, University of Washington, Bothell, WA 98011, USA. [2]College of Forest Resources, [3]Program on the Environment and the School of Marine Affairs, University of Washington, Seattle, WA 98195, USA. [4]Interdisciplinary Arts and Sciences Program, University of Washington, Tacoma, WA 98402, USA.

*To whom correspondence should be addressed. E-mail: wgold@u.washington.edu



Published by AAAS

and neighbors of restored sites, ends the class. This event supports links among the university and surrounding communities and gives students an opportunity to communicate their work.

### Context of the Capstone

The capstone experience is intended for students with prior training in basic concepts of restoration ecology, which they may acquire through a junior-level prerequisite course that is open to students from any academic background. Beyond the academic knowledge of ecology, skills in project management, communication, collaboration, and negotiation are honed on the way to project success.

Through the course of a project, students encounter and learn to address a variety of pragmatic challenges that can be found in many restoration projects. A project may require canvassing neighbors, finding and managing volunteers, preparing education materials, posting signs, attending neighborhood meetings, applying for permits and small grants, soliciting materials, or installing public art.

Capstone students apply scholarly understanding developed in other classes to solve project problems. They must also take the critical step of conveying that knowledge and its application to team members and the public. They learn to communicate the scientific basis of their restoration concept to clients, peers in various disciplines, and the public.

### Our Experiences and Outcomes

Over the past 6 years, 155 students from all three UW campuses have participated in the UW-REN capstone. Of these, 95 have also pursued and been awarded the Restoration Ecology Certificate, which requires additional coursework. The interest in applying academic knowledge to restoring damaged ecosystems occurs across many university departments (*6*, *10*), although the natural sciences dominate our student population (table S3). UW-REN began with courses and faculty from natural science departments. Students from these programs quickly discover UW-REN offerings and are already inclined toward environmental themes. With the natural science foundation in place, we plan to engage a broader set of students and faculty and to add content and skills from fields such as ethics, anthropology, sociology, environmental history, and policy (*11*). Faculty will be recruited who can help students consider such social and humanistic dimensions.

UW-REN offers a model for collaborative education among science and nonscience students. We feel it has been effective in (i) fostering their ability to apply scientific understanding to practical problems in the field, (ii) tackling challenges in a multidisciplinary context and applying their knowledge in a framework with other disciplines, and (iii) developing their abilities to communicate broadly. Through this process, we have watched students come to understand that science is necessary but not sufficient for successful restoration (*6*) and that success depends on effective communication and cross-cutting alignment of values of the project team, clients, and the surrounding community (*12*). Lessons learned in working with community clients have improved our ability to prepare both students and clients for the capstone (*13*).

Our experience has not been without institutional challenges. Although we have had some success, we face repeated challenges in developing ongoing support in terms of funds, faculty time, and resources (*14*).

Students have appreciated how the capstone process reflects the stresses of team-based problem-solving for clients. As one student explained, "The varied perspectives and priorities of our team and clients challenged us to find common ground. As a result, we learned to collaborate in order to develop creative objectives and solutions to accomplish our hybrid goals. The UW-REN Capstone series is a uniquely holistic way of learning that combines the teaching and learning of the theory, application, and social aspects of science." The experience of addressing real-world problems helps students with future challenges (SOM text). Faculty benefit from new teaching and research collaborations. Finally, capstone projects forge rich bonds between the university and community, and provide exemplars of college courses that deal effectively with urgent societal issues and produce workable solutions (*15*). That the community also finds these projects valuable is reflected in our repeat customer rate: More than 40% of our capstone projects have been with clients returning after previous projects (SOM text).

We believe this capstone approach shows how research universities can become engaged with surrounding human and ecological communities while providing students with real-world opportunities to use their education.

**INTERDISCIPLINARY CONTRIBUTIONS TO PROJECT GOALS**

| | Goals | | |
|---|---|---|---|
| | **Maintaining biodiversity** | **Controlling erosion** | **Enhancing sustainability** |
| **Fields contributing disciplinary knowledge** | Conservation biology; ecology; landscape architecture | Civil engineering; botany; geology; soil science | Ecology; education; business; social sciences |
| **Relevant literature on basic science** | Habitat and species diversity; ecosystem stability | Soil properties; slope stability | Succession; nonequilibrium coexistence; assembly rules |
| **Relevant literature on real-world applications** | Wildlife and bird habitat restoration | Erosion control and stabilizing slopes | Successional trajectories |
| **Examples of solutions** | Woody debris addition; microtopography; nest boxes | Bioengineering; wood structure installation | Maintenance; monitoring; stewardship |

**Creating solutions through scientific knowledge.** Students call upon knowledge from the basic and applied literature of a variety of disciplines as they craft solutions to the project's challenges. Table S2 provides literature references and more examples.

### References and Notes


1. National Research Council, *Facilitating Interdisciplinary Research* (National Academies Press, Washington, DC, 2005).
2. M. Hall, *Earth Repair: A Transatlantic History of Environmental Restoration* (Univ. of Virginia Press, Charlottesville, VA, 2005).
3. A. D. Bradshaw, M. J. Chadwick, *The Restoration of Land* (Blackwell, Oxford, 1980).
4. Society for Ecological Restoration, *The SER Primer on Ecological Restoration* (Society for Ecological Restoration Science and Policy Working Group, Tucson, AZ, 2002); available online (www.ser.org).
5. B. Lavendel, *Ecol. Restor*. **17**, 120 (1999).
6. E. Higgs, *Restor. Ecol*. **13**, 159 (2005).
7. Programs in restoration ecology also exist at the University of Wisconsin, Arizona State University, and the University of Victoria, among others.
8. UW-REN (http://depts.washington.edu/uwren) was established with a Tools for Transformation grant from the UW Provost's Office. Cheryl Greengrove (UW Tacoma), Johnny Palka and Sarah Reichard (UW Seattle), and Dan Jaffe (UW Bothell) made early contributions.
9. Disciplines represented among students include biology, landscape architecture, fisheries, engineering, art, geology, urban studies, and education. Current capstone faculty represent various natural sciences.
10. Faculty from at least 12 departments at UW are involved in restoration-related research or teaching.
11. Increased faculty involvement from other disciplines and targeted advertisement should make UW-REN offerings more visible and appealing to students outside the natural sciences. Students already coming from these fields often indicate they did not know these courses existed or were accessible to those outside the natural sciences.
12. P. McManus, *Aust. Geogr.* **37**, 57 (2006).
13. Difficulties with clients include their reluctance to grant students creative independence, inability to grasp the academic nature and requirements of UW-REN projects, and irregularity of client participation. Improved communication has minimized these issues.
14. Support for capstone teaching assistants and some faculty time remain the greatest difficulties. Program courses are now integrated with existing curricula, most faculty are supported by home departments, and administration is coordinated by UW Program on the Environment.
15. The Society for Ecological Restoration International recognized UW-REN's community service contributions with the 2004 John Reiger award.


**Supporting Online Material**
www.sciencemag.org/cgi/content/full/312/5782/1880/DC1

10.1126/science.1128088

CELL SIGNALING

# $H_2O_2$, a Necessary Evil for Cell Signaling

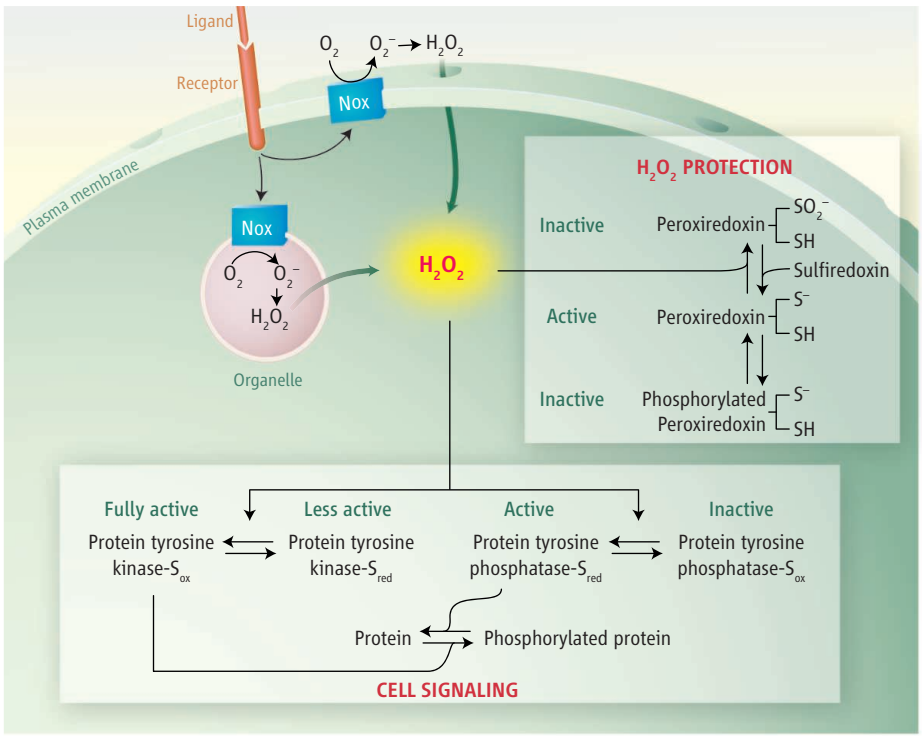
Once considered lethal to cells, reactive oxygen species are now known to be involved in redox signaling pathways that may contribute to normal cell function as well as disease progression.

**Sue Goo Rhee**

For many years, hydrogen peroxide ($H_2O_2$) was viewed as the inevitable but unwanted by-product of an aerobic existence. Given the damage inflicted by $H_2O_2$, it was assumed that the faster the elimination of this toxic waste, the better for the cell. However, as highlighted in recent forums (*1*, *2*), we now know that mammalian cells produce $H_2O_2$ to mediate diverse physiological responses such as cell proliferation, differentiation, and migration (*3*, *4*). This has led to implications of cellular "redox" signaling in regulating normal processes and disease progression, including angiogenesis, oxidative stress and aging, and cancer. This changing view of $H_2O_2$ has partly evolved from a clearer understanding of redox chemistry as it affects biology—that is, cellular signaling that is linked to reductive-oxidative–based mechanisms. As the components and mechanisms involved in performing cellular redox chemistry become better defined, new areas of research are emerging as to how the cells spatially and temporally channel $H_2O_2$ into specific signaling pathways to achieve desired cellular outcomes.

$H_2O_2$ production has been studied most extensively in neutrophils. These immune cells defend a host against infections by engulfing and killing foreign microorganisms. The system relies on Nox [the NADPH (reduced form of nicotinamide adenine dinucleotide phosphate) oxidase complex], which generates millimolar quantities of $H_2O_2$ within the safe confines of an organelle [phagosome (see the figure)] for the purpose of microbial killing. In the classical phagocyte paradigm, stimulation of neutrophils by invading microoganisms leads to assembly at the plasma membrane of an active Nox complex, which comprises a catalytic subunit—the integral membrane protein gp91 Phox—and regulatory proteins including the small guanosine triphosphatase Rac. This complex releases the reactive oxygen species superoxide (the free radical anion $O_2^-$) into the phagosome, and superoxide dismutation yields another reactive oxygen species, $H_2O_2$.

We do know that in nonphagocytic cells, $H_2O_2$ affects numerous intracellular signaling pathways. Nonphagocytic cells express gp91 Phox and its homologs (*5*), and these proteins are the major source of $H_2O_2$ in cells stimulated with various growth factors and cytokines including platelet-derived growth factor (PDGF), epidermal growth factor (EGF), insulin, tumor necrosis factor–α (TNFα), and interleukin-1 (IL-1) (*3*, *4*). However, the coupling of receptor activation to Nox activation in nonphagocytic cells still remains poorly understood.




**$H_2O_2$ production, protection, and signaling actions**. Activation of various cell surface receptors activates Nox situated either in the plasma membrane or in the membrane of organelles such as endosomes to produce $H_2O_2$. To function as an intracellular signaling molecule, $H_2O_2$ must be imported into the cytosol. Cytosolic $H_2O_2$ enhances protein tyrosine phosphorylation by inactivating protein tyrosine phosphatases while activating protein tyrosine kinases. Transient protection of the $H_2O_2$ signal from abundant cytosolic peroxiredoxin appears to result from the reversible inactivation of these enzymes through either hyperoxidation or phosphorylation.

We are also trying to understand the mechanisms by which $H_2O_2$ can modify the activity of key signaling proteins. Biological redox reactions catalyzed by $H_2O_2$ typically involve the oxidation of cysteine residues on proteins, which may affect protein function. Phosphorylation of tyrosine residues in proteins is governed by the opposing activities of protein tyrosine phosphatases and protein tyrosine kinases. The protein tyrosine phosphatase family features a common Cys-X-X-X-X-X-Arg active-site motif (where X = any amino acid). As a result of the invariant arginine, the conserved catalytic cysteine possesses a low p$K_a$ (where $K_a$ is the acid dissociation constant) and exists as a thiolate anion with enhanced susceptibility to oxidation by $H_2O_2$. Oxidation of the essential cysteine abolishes phosphatase activity and can be reversed by cellular thiols. Reversible inactivation of several different protein tyrosine phosphatases has been demonstrated in relevant cell types stimulated with PDGF, EGF, insulin, extracellular matrix molecules, and B cell receptor ligands (*3*, *6*). Oxidative inactivation of these phosphatases and increased tyrosine phosphorylation of target proteins were found to be dependent on $H_2O_2$ production. Moreover, in TNF-α–stimulated cells, the resulting $H_2O_2$ that is generated inactivates mitogen-activated protein kinase phosphatases. This in turn results in sustained activation of c-Jun N-terminal kinase, a subfamily of the mitogen-activated protein kinases that elicits specific cellular responses.

$H_2O_2$ also appears to promote tyrosine phosphorylation by activating protein tyrosine kinases. For example, upon cell attachment to extracellular matrix and associated generation of $H_2O_2$, the tyrosine kinase Src becomes oxidized at two cysteine residues and thus becomes acti-

The author is at the Institute of Molecular Life Science and Technology, Ewha Women's University, Seoul 120-750, South Korea. E-mail: rheesg@ewha.ac.kr





Published by AAAS

vated (*7*). Moreover, antioxidant treatment of cells that express an oncogenic form of Src (v-Src), or mutation of the oxidation-sensitive cysteine residues of v-Src, reduces the potency of v-Src to transform cells. This redox-dependent activation of Src occurs alongside dephosphorylation of a carboxyl-terminal tyrosine, a modification that is needed to activate Src.

For $H_2O_2$ to serve as a signal—through modification of signaling proteins—its concentration must increase rapidly above a certain threshold. How can this occur in the presence of antioxidant enzymes such as catalase, glutathione perioxidase, and peroxiredoxin? Whereas catalase is confined to the peroxisome, several peroxiredoxin isoforms are abundant in the cytosol. Therefore, $H_2O_2$ must be protected from destruction by peroxiredoxin in selected contexts. Indeed, multiple protective mechanisms of this type are being uncovered. During catalysis of $H_2O_2$ reduction, the active-site residue, Cys-SH, of peroxiredoxin occasionally reacts with two molecules of $H_2O_2$, and thus becomes hyperoxidized to Cys-SOOH. Consequently, peroxiredoxins are inactivated (*8*). This inactivation, which can be reversed by sulfiredoxin, an adenosine triphosphate–dependent enzyme, may represent a built-in mechanism to prevent damping of the $H_2O_2$ signal. Prokaryotes do not express sulfiredoxin and their peroxiredoxins are resistant to hyperoxidation. Thus, this regulatory mode appears unique to eukaryotes. Peroxiredoxins are also reversibly inactivated upon phosphorylation by cyclin B–dependent kinase during mitosis (*9*).

Given the toxicity of $H_2O_2$, spatial and temporal regulatory strategies must exist to ensure that Nox activation occurs only where needed and that the $H_2O_2$ signal is terminated in a timely fashion. Recent work on cells stimulated with TNF-α. suggests that Nox proteins are assembled in specific subcellular compartments within membranes such as lipid rafts (*10*). Localized Nox assembly also occurs at focal complexes, points of contact between a moving cell and the extracellular matrix, in response to migratory stimuli (*11*). The relevant oxidation targets that are presumably enriched in these microenvironments remain to be identified.

Despite the increasingly sophisticated molecular descriptions of $H_2O_2$ action, disturbingly little is understood about how $H_2O_2$ is actually delivered to the cytosol. The classical neutrophil studies demonstrate that Nox releases $H_2O_2$ into the phagosome, which is topologically equivalent to the extracellular space. How, then, does $H_2O_2$ modulate intracellular signaling? In one scenario, Nox situated at the plasma membrane releases $H_2O_2$ into the extracellular space as an autocrine factor to be imported into the cell. Alternatively, Nox proteins assembled at organelle membranes discharge $H_2O_2$ into the lumenal space. For example, binding of IL-1 to its receptor in the plasma membrane triggers Rac-mediated Nox association with the IL-1 receptor and endocytosis (internalization) of the receptor complex (*12*). This results in superoxide production and conversion into $H_2O_2$ in the lumen of the endosome. In addition, Nox isoforms and their regulatory subunits have been detected in other cell organelles including the endoplasmic reticulum and nucleus.

Regardless of whether the Nox complex is activated at the cell surface or within an organelle, the resultant $H_2O_2$ must traverse the lipid bilayer to access the cytosol, where most if not all of its target proteins exist. Although $H_2O_2$ is believed to diffuse freely across membranes, recent studies indicate that some membranes are poorly permeable to $H_2O_2$. Instead, $H_2O_2$ transport might be regulated by changes in membrane lipid composition or by aquaporins (*13*), which are diffusion-facilitating channel proteins for noncharged solutes such as water.

The current picture of $H_2O_2$-based redox regulation of signaling processes is rapidly expanding beyond those issues focused on here. The development of a sensitive and specific probe for $H_2O_2$ that allows quantitative and dynamic assessment in live cells, conspicuously lacking in studies to date, will be a great boon for the study of this misunderstood and maligned molecule.

**References and Notes**


1. Redox Signalling in Human Disease and Ageing, Catholic University, Rome, Italy, 20 to 23 April 2006 (EMBO Conference Series 2006).
2. Thiol-Based Redox Regulation and Signaling, University of New England, Bedford, ME, USA, 18 to 23 June 2006 (Gordon Research Conferences 2006).
3. S. G. Rhee, *Sci. STKE* **2000**, pe1 (2000).
4. M. Sundaresan, *Science* **270**, 296 (1995).
5. J. D. Lambeth *et al.*, *Nat. Rev. Immunol*. **4**, 181 (2004).
6. N. Tonks, *Cell* **121**, 667 (2005).
7. E. Giannoni *et al.*, *Mol. Cell. Biol*. **25**, 6391 (2005).
8. S. G. Rhee *et al.*, *Curr. Opin. Cell Biol*. **17**, 183 (2005).
9. T. S. Chang *et al.*, *J. Biol. Chem.* **277**, 25370 (2002).
10. F. Vilhardt, B. van Deurs, *EMBO J*. **23**, 739 (2004).
11. R. F. Wu *et al.*, *J. Cell Biol*. **171**, 893 (2005).
12. Q. Li *et al.*, *Mol. Cell. Biol*. **26**, 140 (2006).
13. G. B. Bienert *et al.*, *Biochim. Biophys. Acta*, in press.

14. S.G.R.'s research is supported by Grant FPR0502-470 from the Korean Ministry of Science and Technology.

10.1126/science.1130481

MATERIALS SCIENCE

# Seeking Room-Temperature Ferromagnetic Semiconductors

**Koji Ando**

Microelectronic circuits that retain their logic state when the power is off would permit entirely new kinds of computers. Ferromagnetic semiconductors might make this technology possible.

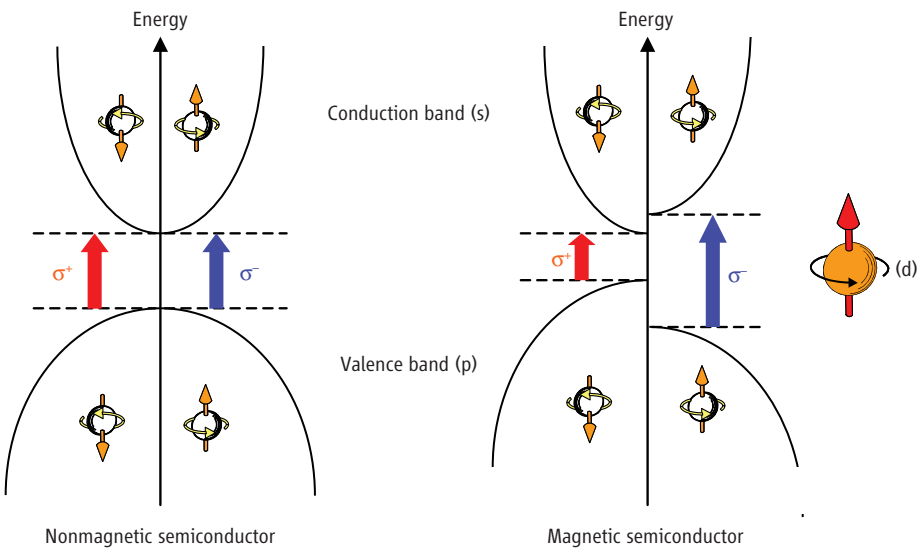
Nonvolatile digital circuits that retain their logic states even when their power sources are rapidly switched on and off would make possible a new type of computer. Although appearing to operate normally, these devices would actually be turned off most of the time, potentially changing the way we use digital devices. Such devices would allow, for example, year-long operation of mobile computers, an enormous number of tiny computers embedded all around us to help our daily lives, and ultra-high-density integrated circuits free from heat generation problems. To make this dream a reality, nonvolatile transistors are needed, but unfortunately this technology is nonexistent. Because transistors are composed of semiconductor materials, the ideal way to make nonvolatile transistors would be to use semiconductor materials that are intrinsically nonvolatile.

The author is at the National Institute of Advanced Industrial Science and Technology (AIST), Nanoelectronics Research Institute, Umezono 1-1-1, Central 2, Tsukuba, Ibaraki 305-8568, Japan. E-mail: ando-koji@aist.go.jp

Among several physical phenomena that produce nonvolatility, the most enticing is that of ferromagnetic hysteresis. In this effect, the material retains its magnetic state until reversed by a suitable magnetic field. Ferromagnetism has been verified to offer high-speed, unlimited magnetization reversal, so it is perfect for transistor applications. However, the ferromagnetic materials used in digital devices such as hard disks and magnetic random access memory chips—iron, cobalt, and nickel and their alloys—are not semiconductor materials. Hence, there is a continuing search for semiconductor materials that display ferromagnetic properties.

By replacing some of the positive ions of the parent nonmagnetic semiconductors by magnetic ions, one can make ferromagnetic semiconductors such as (In,Mn)As and (Ga,Mn)As (*1*). But their ferromagnetic Curie temperatures ($T_c$)—the temperature at which the ferromagnetism disappears—are 61 K (*2*) and 173 K (*3*), respectively, much lower than room temperature. In 2000, Dietl *et al.* (*4*) used a simple theory to estimate the $T_c$ of

ferromagnetic semiconductors, and they predicted that room-temperature ferromagnetic semiconductors might be created by substituting manganese ions in wide–band gap semiconductors such as GaN and ZnO. Reports of GaN- and ZnO-based room-temperature ferromagnetic semiconductors (*5*, *6*) soon followed. A Curie temperature as high as 940 K was reported for GaN with less than 10% manganese (*5*). Before long, reports began appearing of room-temperature ferromagnetic semiconductors extended to materials based on other useful oxide insulators such as $TiO_2$ (*7*). Eventually there appeared a report of room-temperature ferromagnetism of pure $HfO_2$ (*8*), which contains no magnetic ions at all. In the physics world it was long held that ferromagnetism at room temperature could only be achieved with materials containing very high concentrations of magnetic ions, so these reports came as a big surprise.




**Magnetism in semiconductors. (Left)** In normal, nonmagnetic semiconductors, electronic energy does not depend on the spin direction. It is not possible to distinguish between spin-up and spin-down electrons. **(Right)** In magnetic semiconductors, the d electrons of magnetic ions influence the s and p electrons, and the conduction band and valence band are split depending on the spin direction (Zeeman splitting). This spin-polarized semiconductor band structure alters the absorption of clockwise-polarized ($\sigma^+$) and counterclockwise-polarized ($\sigma^-$) light (the MCD effect). For this reason, a magnetic semiconductor should display an MCD spectrum that reflects the band structure of the parent semiconductor.

Are we close to realizing the dream of nonvolatile transistors? Half a decade has already passed since the first report of the high-$T_c$ ferromagnetic semiconductor, but debate on the nature of the observed ferromagnetic signals still continues to rage (*9*–*12*). Now that the excitement has passed, efforts are increasingly shifting to the task of thinking calmly about ways to evaluate ferromagnetic semiconductors.

The concentration of magnetic ions that can be introduced into a parent semiconductor is at most 10%, even with the use of a nonequilibrium film growth method. For this reason, the magnetization displayed by thin-film specimens of what is claimed to be a "ferromagnetic semiconductor" is typically only $10^{-5}$ emu. Still, this level of weak magnetization can be easily detected by extremely high-sensitivity SQUIDs (superconducting quantum interference devices). However, careful attention should be paid to the fact that even tiny amounts of iron—as little as 1 part per 2000 of the typical sample volume used in SQUID measurements—can generate these kinds of magnetic signals. A recent report (*13*) cautioned that nonmagnetic $HfO_2$ thin films generate clear, ferromagnetic signals after contact with stainless steel tweezers. Because new, unexpected ferromagnetic materials can be synthesized in the nonequilibrium growth method, the possible presence of very small amounts of an impurity phase is a critical problem that cannot be avoided in research on ferromagnetic semiconductors. Crystallographic evaluation methods, such as x-ray diffraction and transmission electron microscopy, are often used in claims of the absence of impurities, but their detection sensitivity is extremely low relative to that of SQUID, and so they cannot be taken to provide conclusive evidence.

To clear up this confusion, it is useful to consider the essential character of a magnetic semiconductor. If magnetic properties and semiconductor properties existed independently of each other, magnetic semiconductors would have absolutely no value. It is the mutual interaction between magnetic properties (supported by the d electrons of the magnetic ions) and the semiconductor properties (supported by the s and p electrons) that is the essence of a magnetic semiconductor. With this s, p–d exchange interaction, one can control the magnetization by the electrical field (*14*) or control the semiconductor optical characteristics by the magnetic field (*15*), for example. Because of the s, p–d exchange interaction, the energies of the s and p electrons in a magnetic semiconductor become dependent on their spin state [up or down (see the figure)]. Therefore, confirmation of a spin-polarized semiconductor band is the litmus test of the existence of a magnetic semiconductor.

As a result of the quantum mechanical selection rules for optical absorption, the spin polarization of a semiconductor band is directly connected with the appearance of magnetic circular dichroism (MCD) (*15*). MCD is an effect in which clockwise-polarized and counterclockwise-polarized light are absorbed differently. The important point here is that, whereas all magnetic materials can produce an MCD signal, a magnetic semiconductor should not only have a strong MCD signal but should also display an MCD spectral shape that reflects the band structure of the parent semiconductor (*15*). If the MCD spectral shape is different from what is expected, it is possible that the detected MCD signal is being produced by a magnetic material other than the expected magnetic semiconductor. When the magnetic field dependence of the MCD signal of a magnetic semiconductor behaves ferromagnetically, we can conclude that we have a ferromagnetic semiconductor.

Up to now, studies of the spin-polarized semiconductor band structures of materials claimed to be "ferromagnetic semiconductors" have been limited despite their importance. In the cases of GaN:Mn, GaAs:Cr, and ZnO:Ni—all of which produced ferromagnetic SQUID signals—no magnetic MCD signals were observed (*16*). It is very likely that some material other than a magnetic semiconductor was responsible for the observed ferromagnetic SQUID signal. The MCD spectral shape displayed by a ferromagnetic ZnO:Co was very different from that of paramagnetic (Zn,Co)O (*16*). In the case of $TiO_2$:Co, no clear correspondence between the shape of the ferromagnetic MCD spectra (*17*) and the expected band structure has been established. On the other hand, MCD spectral analysis has verified that (In,Mn)As (*18*), (Ga,Mn)As (*19*), and (Zn,Cr)Te (*20*) are intrinsic ferromagnetic semiconductors.

To make one's way in a wilderness, a compass is necessary. In the quest for ferromagnetic semiconductors, that compass is MCD analysis and theoretical materials design. Recently, the number of experimental MCD studies has been increasing, but hardly any theoretical work has been done on interpreting the spectral shapes in MCD. On the other hand, there have been substantial improvements in theories for predicting the $T_c$ of ferromagnetic semiconductors (*21*, *22*). It seems that the wide–band gap semiconductors do not necessarily have an advantage in achieving high $T_c$. An alternative would be to look more

closely at materials having a longer exchange interaction distance between magnetic ions and a higher concentration of magnetic ions. All of the (In,Mn)As, (Ga,Mn)As, and (Zn,Cr)Te materials belong to this category. Recently, the $T_c$ of (Ga,Mn)As has been increasing from an initial value of 110 K (*1*) up to 173 K (*3*) by improving the growth procedures, and that of (Zn,Cr)Te has reached 300 K (*20*).

Many different characteristics are considered when searching for the ferromagnetic semiconductors needed for nonvolatile transistors, including high $T_c$, high carrier mobility, and intrinsic insulation conditions for carrier doping. To find our goal, it seems we must continue wandering, with the compass of MCD in one hand.


**References**

1. H. Ohno, *Science* **281**, 951 (1998).
2. H. Ohno, *J. Cryst. Growth* **251**, 285 (2003).
3. T. Jungwirth *et al.*, *Phys. Rev. B* **72**, 165204 (2005).
4. T. Dietl, H. Ohno, F. Matsukura, J. Cibert, D. Ferrand, *Science* **287**, 1019 (2000).
5. S. Sonada *et al.*, *J. Cryst. Growth* **237–239**, 1358 (2002).
6. K. Ueda, H. Tabata, T. Kawai, *Appl. Phys. Lett.* **79**, 988 (2001).
7. Y. Matsumoto *et al.*, *Science* **291**, 854 (2001); published online 11 January 2001 (10.1126/science.1056186).
8. M. Venkatesan, C. B. Fitzgerald, J. M. D. Coey, *Nature* **430**, 630 (2004).
9. K. Ando, *Appl. Phys. Lett.* **82**, 100 (2003).
10. D. C. Kundaliya *et al.*, *Nat. Mater.* **3**, 709 (2004).
11. S. R. Shinde *et al.*, *Phys. Rev. Lett.* **92**, 166601 (2004).
12. C. Liu, F. Yun, H. Morkoc, *J. Mater. Sci.* **16**, 555 (2005).
13. D. W. Abraham, M. M. Frank, S. Guha, *Appl. Phys. Lett.* **87**, 252502 (2005).
14. H. Ohno *et al.*, *Nature* **408**, 944 (2000).
15. K. Ando, in *Magneto-Optics*, S. Sugano, N. Kojima, Eds., vol. 128 of *Springer-Verlag Series in Solid-State Science* (Springer, Berlin, 2000), pp. 211–244.
16. K. Ando, H. Saito, V. Zayets, M. C. Debnath, *J. Phys. Condens. Matter* **16**, S5541 (2004).
17. H. Toyosaki, T. Fukumura, Y. Yamada, M. Kawasaki, *Appl. Phys. Lett.* **86**, 182503 (2005).
18. K. Ando, H. Munekata, *J. Magn. Magn. Mater.* **272–276**, 2004 (2004).
19. K. Ando, T. Hayashi, M. Tanaka, A. Twardowski, *J. Appl. Phys.* **83**, 6548 (1998).
20. H. Saito, V. Zayets, S. Yamagata, K. Ando, *Phys. Rev. Lett.* **90**, 207202 (2003).
21. K. Sato *et al.*, *Phys. Rev. B* **70**, 201202R (2004).
22. L. Bergqvist *et al.*, *Phys. Rev. B* **72**, 195210 (2005).


10.1126/science.1125461

CHEMISTRY

# Targeting Specific C-H Bonds for Oxidation

Enzymes use two main strategies—size/shape selectivity and substrate orientation—to achieve selective oxidation. Chemists have now developed a simple catalyst that mimics substrate orientation.

**Rubén Mas-Ballesté and Lawrence Que Jr.**

A main challenge in chemistry is the design of catalysts that can carry out a desired chemical transformation only upon chosen specific targets. These specific targets can be one molecule in a mixture of chemical substances or—perhaps even more difficult—one specific site in a molecule with several reactive sites. On page 1941 of this issue, Das *et al.* (*1*) report an innovative approach to the design of a bio-inspired catalyst for the highly selective oxidation of CH bonds.

Nature uses different strategies to achieve selective oxidations. For instance, the enzyme methane monooxygenase (MMO) selectively oxidizes methane to methanol, even in the presence of other hydrocarbons with weaker C-H bonds (*2*). This specificity arises from a sieve-like phenomenon that hampers the access of molecules bigger than methane to the active site of the enzyme (see the figure, left panel) (*3*, *4*). Besides size (and shape), enzymes can confer specificity by means of noncovalent interactions that orient the substrate in a particular manner (see the figure, right panel). Fatty acid desaturases use a functional group within the enzyme active site to anchor the carboxylate end of a fatty acid and achieve dehydrogenation at two specific carbon atoms (*5*).

Finding inspiration in nature, chemists have designed synthetic systems to emulate the specificity of biological oxidations. An initial approach reported by Breslow used steroid substrates with a dangling, covalently attached photosensitive group to carry out light-activated remote oxidations at particular sites of the substrate (*6*). In the same vein, Stuk *et al.* appended a substrate to a manganese catalyst to achieve selective oxidation at specific sites of the substrate (*7*). These examples demonstrated the possibility of selective oxidations in a biomimetic system, but the covalent binding of substrate to the oxidant constrained the chemistry to a maximum of a single turnover, such that only stoichiometric (rather than catalytic) oxidation could be achieved.

Subsequent efforts have incorporated into the catalyst design moieties that can recognize substrates through noncovalent interactions. An approach extensively explored in the past used a metalloporphyrin as the catalytic center, with modifications at its periphery to confer substrate selectivity.

For example, Bhyrappa *et al.* attached dendrimers to the metalloporphyrin to favor the oxidation of less sterically hindered linear alkanes over cyclic alkanes (*8*); this approach emulates the size-recognition strategy of MMO. Mimicking the anchoring strategy of fatty acid desaturases, Groves and Neumann (*9*) embedded a metallo-


**Size/shape recognition versus substrate orientation.** Schematic representation of the two main strategies used by enzymes to achieve selectivity.

CREDIT: P. HUEY/*SCIENCE*

The authors are in the Department of Chemistry and Center for Metals in Biocatalysis, University of Minnesota, Minneapolis, MN 55455, USA. E-mail: que@chem.umn.edu



*Published by AAAS*

porphyrin catalyst into a phospholipid bilayer that selected for hydrophobic substrates and oriented them in such a way as to attain regioselective epoxidation of sterols and polyunsaturated fatty acids. Combining both size-recognition and substrate-orientation design elements (see the figure), Breslow *et al.* attached β-cyclodextrin groups to the metalloporphyrin periphery to bind substrates with appropriately sized and suitably positioned hydrophobic groups (*10*). This binding step served to bring only the target C–H bond into close proximity to the metal center; hundreds to thousands of turnovers could be achieved (*11*).

Das *et al.* now report the design and synthesis of a new bio-inspired catalyst for the highly selective oxidation of C–H bonds. Instead of a metalloporhpyrin, they have used a dinuclear manganese complex as the catalytic center, where the Mn ions are coordinated to a tridentate ligand to which a carboxylic acid group (the tweezers in the right panel) is attached via a suitably rigid framework (the tether in the right panel). This group can interact via hydrogen bonds with the carboxylic group of the substrate, orienting it in such a way that only one oxidizable site is in the correct position to be attacked by the metal center. Following this strategy, not only is a high regioselectivity achieved but also an excellent stereoselectivity for the oxidation of a substrate that can adopt different conformations. The system reported in this issue mimics very efficiently (up to 700 catalytic turnovers have been observed) the strategy used by fatty acid desaturases with a simple and elegant catalyst design.

The relative simplicity of the ligand design used by Das *et al.* allows us to envisage a new horizon of modified ligands. By tuning the nature of the tweezer and the length of the tether, a whole family of catalysts could be tailored to accommodate a wide range of substrates. This work, together with the recent results reported by Breslow and co-workers (*10*, *11*), demonstrates the viability of molecular recognition strategies for catalytic oxidations.


**References**

1. S. Das. C. D. Incarvito, R. H. Crabtree, G. W. Brudvig, *Science* **312**, 1941 (2006).
2. M. Merkx *et al.*, *Angew. Chem. Int. Ed.* **40**, 2782 (2001).
3. B. J. Brazeau, J. D. Lipscomb, *Biochemistry* **39**, 13503 (2000).
4. H. Zheng, J. D. Lipscomb, *Biochemistry* **45**, 1685 (2006).
5. B. G. Fox, K. S. Lyle, C. Rogge, *Acc. Chem. Res.* **37**, 421 (2004).
6. R. Breslow, *Acc. Chem. Res.* **13**, 170 (1980).
7. T. L. Stuk, P. A. Grieco, M. M. Marsh, *J. Org. Chem.* **56**, 2957 (1991).
8. P. Bhyrappa, J. K. Young, J. S. Moore, K. S. Suslick, *J. Am. Chem. Soc.* **118**, 5708 (1987).
9. J. T. Groves, R. Neumann, *J. Am. Chem. Soc.* **109**, 5045 (1987).
10. R. Breslow, Y. Huang, X. Zhang, J. Yang, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 11156 (1997).
11. J. Yang, B. Gabriele, S. Belvedere, Y. Huang, R. Breslow, *J. Org. Chem.* **67**, 5057 (2002).


10.1126/science.1129814

MOLECULAR BIOLOGY

# Accurate RNA Siting and Splicing Gets Help from a DEK-Hand

Mature messenger RNA is prepared in eukaryotes by splicing together the protein-coding sequences of precursors. A chromatin-associated protein proofreads these splice sites in the RNA, ensuring accuracy.

**Tracy L. Kress and Christine Guthrie**

Eukaryotic genomes are very economical, as the information contained in a single genome is usually greater than the sum of its genes. Different proteins can be generated from a single gene [specifically, from the initial precursor messenger RNA (pre-mRNA) that is synthesized from a single gene] through a complex splicing process. By excising different introns and splicing together different exons—the noncoding and coding portions of transcribed pre-mRNA, respectively—a gene can expand its repertoire of mRNA-encoded products. But this relies on the exquisite accuracy of the cellular "splicing machinery" to accomplish this task. On page 1961 of this issue, Menedes Soares *et al.* (*1*) report that accurate identification of splice sites involves a factor that is not part of the known splicing machinery that carries out the cutting and pasting of mRNA, but is associated with chromatin, the material of chromosomes.

Enhanced online at www.sciencemag.org/cgi/content/full/312/5782/1886

The spliceosome is a large and dynamic ribonucleoprotein machine that removes introns from pre-mRNAs (*2*). Understanding how the 5′ and 3′ ends of introns are accurately identified, despite relatively degenerate consensus sequences, remains a fundamental challenge. Errors of even a single nucleotide in the specification of an intron boundary can have catastrophic consequences—for example, by creating a frameshift mutation. This problem is magnified by the widespread existence of alternative splicing, in which a given splice site can be recognized in one biological context and ignored in another. Thus, the spliceosome must have evolved mechanisms to balance precision and plasticity in splice site selection. Extensive genetic analyses in budding yeast have revealed a number of factors that, when altered, degrade fidelity of splice site choice by allowing the use of mutated intron boundaries (*3*–*7*). To date, all of these factors are components of the core splicing machinery. Mendes Soares *et al.* now report that DEK, a known chromatin-associated protein frequently overexpressed in tumor cells (*8*), plays a key role in splice site recognition.

The 3′ ends of introns in metazoa are defined by a composite signal: a polypyrimidine (Py) tract upstream of an AG dinucleotide that is the 3′ splice site (see the figure). These sequences are recognized by the RNA-binding proteins U2AF$^{65}$ (*9*) and U2AF$^{35}$ (*10*–*12*), respectively, which can form a heterodimer. When the Py tract is sufficiently long, the AG sequence is not required until the second of the two chemical steps involved in splicing (hence, the term AG-independent) (*10*, *13*, *14*). In contrast, so-called AG-dependent introns, which have short or interrupted Py tracts *(13)*, require U2AF$^{35}$ binding to promote or stabilize the binding of U2AF$^{65}$ to the weak Py tracts (*10*, *14*, *15*). Earlier work (*10*–*12*, *14)* indicated that U2AF$^{35}$ might be sufficient to confer sequence specificity for the AG dinucleotides. However, Mendes Soares *et al.* now report that purified U2AF$^{35}$ can promote binding of purified U2AF$^{65}$ to artificial RNAs containing a Py tract and either a wild-type 3′-AG or a mutated 3′ splice site, indicating that U2AF$^{35}$ alone is not sufficient to differentiate between the two. Although the authors do not pinpoint differences between the present and published assays for splice site discrimination that might account for this discrepancy, they show that specificity for the correct AG dinucleotide splice site can be restored upon addition of a nuclear extract, made from a human cell line, that is depleted of the U2AF heterodimer. This provides the first evidence that an additional factor or activity is required for maximal 3′ splice site specificity.

To identify this mystery factor, the authors used an ultraviolet light–induced oligonucleotide cross-linking assay to monitor 3′-AG splice site discrimination activity during the purification of

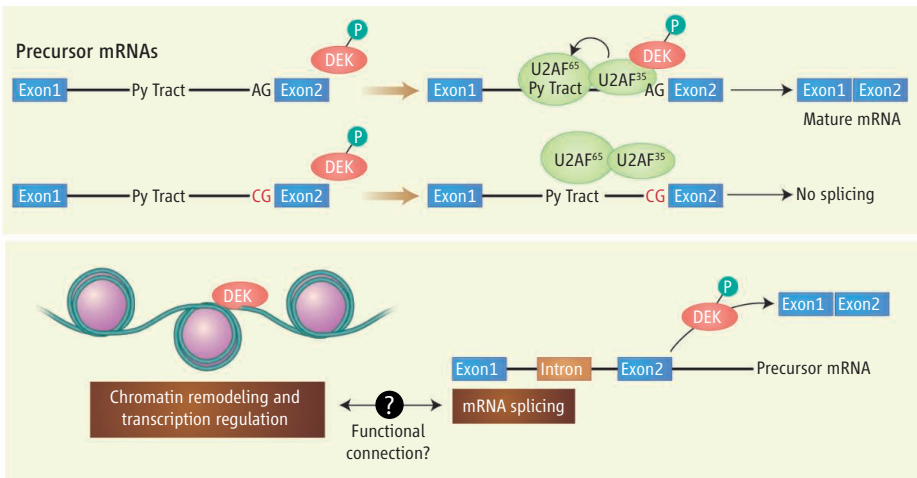
The authors are in the Department of Biochemistry and Biophysics, University of California, San Francisco, CA 94143–0448, USA. E-mail: guthrie@biochem.ucsf.edu

U2AF from nuclear extracts. They isolated a 50-kD protein identified by mass spectrometry as DEK, a protein originally associated with leukemia (*8*). Depletion of DEK from a nuclear extract reduced the ability of endogenous U2AF to discriminate between CG and AG dinucleotides, and this activity was substantially restored by the addition of recombinant DEK protein. Moreover, an interaction between in vitro–synthesized U2AF$^{35}$ and recombinant DEK protein was observed. DEK is known to be phosphorylated in cells; recombinant DEK is not phosphorylated but may become so upon addition to nuclear extract or when added to in vitro–synthesized U2AF$^{35}$. The authors show that phosphorylation of DEK is required for this interaction as mutation of two known serine phosphorylation sites in DEK both abolished DEK phosphorylation and inhibited its interaction with U2AF$^{35}$. When the authors repeated the cross-linking assay with recombinant DEK and purified U2AF heterodimer, maximal splice site discrimination activity required treatment of DEK with nuclear extract (which was depleted of DEK and U2AF through biochemical fractionation). The authors speculate that the extract may contain a kinase required to phosphorylate DEK, and show that recombinant DEK can be phosphorylated upon addition to a nuclear extract. Indeed, the increase in 3′ splice site discrimination correlated with phosphorylation of DEK, whereas the phosphorylation-defective DEK mutants showed decreased discrimination. These results suggest that phosphorylation of DEK promotes its association with U2AF$^{35}$, which in turn enhances AG dinucleotide discrimination by the U2AF heterodimer.

Does DEK function solely to suppress the use of noncanonical 3′ splice sites? Depletion of DEK from nuclear extracts did not lead to splicing at mutant 3′ splice sites in either of two pre-mRNAs tested (one AG-dependent and one AG-independent). Notably, however, these DEK-depleted extracts failed to splice wild-type RNA substrates. This defect could be rescued by the addition of recombinant DEK. Interestingly, removal of the AG-independent intron was blocked specifically at the second chemical step, arguing that in the absence of DEK, spliceosomes could still assemble and carry out first-step chemistry. In this case, DEK's role would be to ensure that an AG dinucleotide has been identified at the 3′ splice site before allowing the chemical reactions to proceed, as has previously been shown for U2AF$^{35}$ (*10*, *14*, *15*). To test this hypothesis directly, the authors assayed an early step in spliceosome assembly—the binding of U2 small nuclear ribonucleoprotein (snRNP) to pre-mRNA or oligonucleotides containing a 3′ splice site. When a 3′-AG–containing RNA was used, U2 snRNP binding was unaffected by the absence of DEK. However, when snRNP binding was assayed in the presence of RNA competitors, an AG-containing RNA could outcompete a CG-containing RNA, as long as phosphorylatable DEK was present. By this criterion, DEK might function early in spliceosome assembly, to enhance the discrimination between wild-type and noncanonical 3′ splice sites.

The authors conclude that DEK acts to proofread the 3′ splice site. How does this fit into the current view of fidelity mechanisms? Proofreading factors identified to date play important roles in the splicing of wild-type introns (*3*). That is, like DEK, their function is not restricted to fidelity per se. As components of the splicing machinery, they influence the rate of interconversion between alternative conformations of the spliceosome during the chemical steps of splicing, likely underlying the extent to which aberrant splice sites are used (*3*, *6*). These conformational changes are often coupled to an irreversible step, such as the hydrolysis of ATP (adenosine triphosphate) by DEAD-box RNA-dependent ATPases, in a kinetic proofreading scheme (*7*). As the authors point out, it will be exciting to investigate the possibility that DEK is involved in kinetic proofreading by perhaps modulating the activity of an ATPase. Interestingly, recent results from yeast implicate Prp22, a DEAD-box RNA-dependent ATPase splicing factor, in the rejection of mutated 3′ splice sites immediately before exon ligation. By slowing an ATP-dependent "timer," mutations that reduce the ATPase activity of Prp22 can allow aberrant introns to be spliced (*6*).

Perhaps the most provocative questions raised by this work involve the potential link between splicing and chromatin remodeling/transcriptional regulation, given that DEK is known to be chromatin-associated. Intriguingly, phosphorylation of DEK, which is required for splicing activation, reduces its affinity for DNA (*16*). Conceivably, the phosphorylation state of DEK acts in a switchlike fashion to alter the association of DEK between proteins involved in chromatin remodeling and transcriptional regulation and proteins involved in RNA splicing. It will also be important to determine whether DEK plays a role in alternative splicing, and if so, how phosphorylation may influence this activity. In addition to elucidating the role of DEK in the fidelity of splice site choice, such information could help explain how the misregulation of DEK might underlie disease.




**RNA splice site discrimination requires DEK. (Top)** Phosphorylated (P) DEK, which interacts with the splicing factor U2AF$^{35}$, allows accurate selection of the 3′-AG dinucleotide sequence and allows U2AF$^{65}$ to rapidly assemble on the precursor mRNA. This may outcompete a precursor mRNA that contains a mutated 3′ splice site. **(Bottom)** It is conceivable that DEK, which associates with chromatin, may coordinate chromatin remodeling/transcriptional regulation and mRNA splicing. Phosphorylation may operate as a switch for DEK between these two functions.

CREDIT: P. HUEY/*SCIENCE*


**References**

1. L. M. Mendes Soares, K. Zanier, C. Mackereth, M. Sattler, J. Valcárcel, *Science* **312**, 1961 (2006).
2. C. L. Will, R. Luhrmann, in *The RNA World*, T. R. Cech, J. F. Atkins, Eds. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, ed. 3, 2006), pp. 369–400.
3. M. M. Konarska, C. C. Query, *Genes Dev.* **19**, 2255 (2005).
4. C. C. Query, M. M. Konarska, *Mol. Cell* **14**, 343 (2004).
5. T. Villa, C. Guthrie, *Genes Dev.* **19**, 1894 (2005).
6. R. M. Mayas, H. Maita, J. P. Staley, *Nat. Struct. Mol. Biol.* **13**, 482 (2006).
7. S. M. Burgess, C. Guthrie, *Cell* **73**, 1377 (1993).
8. T. Waldmann, I. Scholten, F. Kappes, H. G. Hu, R. Knippers, *Gene* **343**, 1 (2004).
9. P. D. Zamore, J. G. Patton, M. R. Green, *Nature* **355**, 609 (1992).
10. S. Wu, C. M. Romfo, T. W. Nilsen, M. R. Green, *Nature* **402**, 832 (1999).
11. D. A. Zorio, T. Blumenthal, *Nature* **402**, 835 (1999).
12. L. Merendino *et al.*, *Nature* **402**, 838 (1999).
13. R. Reed, *Genes Dev.* **3**, 2113 (1989).
14. S. Guth *et al.*, *Mol. Cell. Biol.* **21**, 7673 (2001).
15. P. Zuo, T. Maniatis, *Genes Dev.* **10**, 1356 (1996).
16. F. Kappes *et al.*, *Mol. Cell. Biol.* **24**, 6011 (2004).


10.1126/science.1130324



*Published by AAAS*

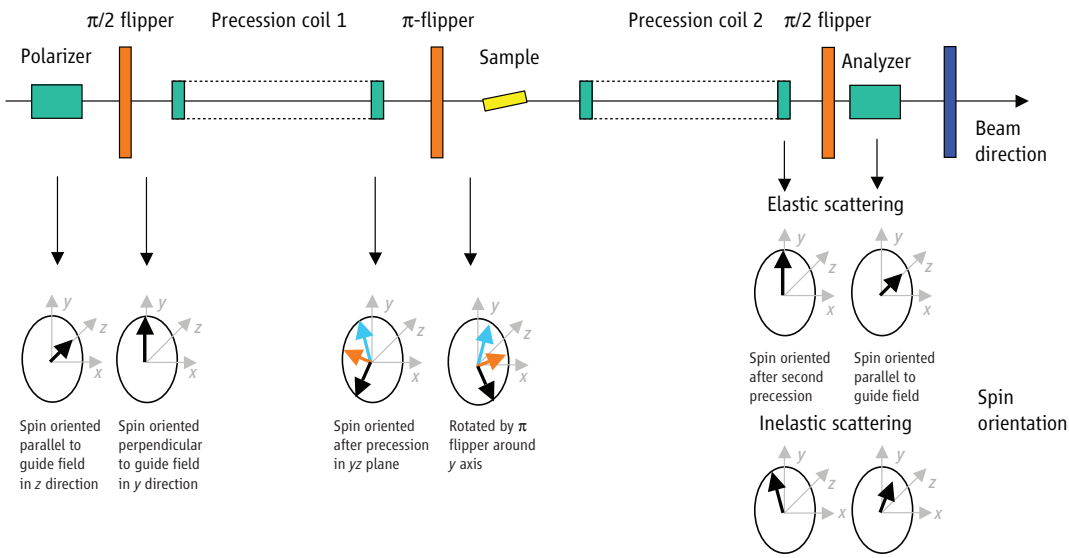
APPLIED PHYSICS

# The Neutron Spin-Echo Technique at Full Strength

Improvements in neutron scattering techniques allow precise measurement of dynamics in magnetically ordered condensed matter materials.

**Joël Mesot**

The dynamics of individual particles such as atoms, electrons, and their spins changes in a dramatic manner when these particles form a periodic solid. Although the energy $E$ of a free particle of mass $M$ relates to its velocity $\mathbf{V}$ (or momentum $\mathbf{P} = M\mathbf{V}$) through the well-known relation $E = M\mathbf{V}^2/2$ (or $E = \mathbf{P}^2/2M$), in a crystal this relation is replaced by the more general relation $E(\mathbf{P})$, which is called a dispersion relation. This dispersion relation can take very intricate forms that encode the periodic environment as well as the interactions between the elementary building blocks of the solid. Since the work of Landau on Fermi liquids (*1*), this dispersion relation has been recognized as a basic and essential concept in understanding weakly interacting quasiparticles, such as, among others, phonons responsible for lattice vibrations, electrons responsible for electrical conductivity, and magnons responsible for spin waves. Now, as reported on page 1926 of this issue, Bayrakci *et al.* (*2*) have taken the measurement of the properties of such quasiparticles to a new level of sophistication.




**Spin gymnastics.** Measurement of the neutron spin polarization after interaction with a sample provides information about dynamical processes occurring in the investigated materials.

The validity of Landau's approach has been directly verified for electrons by means of angle-resolved photoelectron spectroscopy (ARPES) and for phonons with both inelastic neutron and x-ray scattering techniques. Magnon dispersions can only be measured by inelastic neutron scattering, however, which therefore represents the technique of choice to validate or invalidate our current theoretical understanding of magnetic phenomena in condensed matter physics.

True noninteracting quasiparticles can exist only at both zero temperature and at the minimum of the dispersion relation, where no decay can occur. These noninteracting particles are characterized by infinite lifetime, which is related, through the quantum uncertainty principle, to a zero linewidth in an energy spectrum (energy width). As one moves away from the energy minimum, scattering between quasiparticles (phonon-phonon, electron-electron, magnon-magnon, or mixed interactions) becomes possible and, as a consequence, their lifetimes diminish and their energy spectral widths increase.

Experimentally, it is much more difficult to determine lifetime effects than dispersion relations. This is mainly a result of the fact that in most conventional scattering experiments, such as those that use neutron three-axis spectrometers (TAS), the energy resolution is inversely proportional to the beam intensity. Ideally, one would like to have a method that allows one to disentangle the resolution from the intensity. Such a technique exists for neutron scattering and is based on the spin-echo effect, where one uses the "independent" spin degree of freedom of the neutron (instead of its momentum or wavelength) to achieve high resolution and high intensity (see the first figure). The neutron spin-echo technique was invented by Mezei in 1972 (*3*) and was first realized at the high-flux reactor of the Institute Laue Langevin (ILL) in Grenoble, France, in 1978. As the illustration shows, both a polarizer and a spin flipper control and guide the neutron spins into a first precession coil. Because the number of precessions depends upon the velocity (indicated by colored arrows) of each neutron, the polarization is lost at the sample position. After reversal of their spin around the $y$ axis (by means of a π-flipper), the neutrons pass through a second precession coil and a spin analyzer, to be finally detected. Although the original polarization is fully recovered (the spin echo) for elastically scattered neutrons, neutrons scattered inelastically by an excitation will result in a tilted polarization. The magnitude of this tilt is a measure of dynamical processes in the sample.

The main applications of this technique have been very-high-resolution quasielastic measurements of mesoscopic systems in the fields of soft matter, biology, and glasses. It turned out that traditional spin-echo is, for various reasons, not well adapted for measurements of dispersion curves away from their minima. To overcome this limitation, it was suggested that traditional TAS could be combined with a modified version of the spin-echo technique. After this proposal, several devices were realized at the Hahn-Meitner-Institute in Berlin (*4*), ILL (*5*), and the Forschungsreaktor München (FRM) in Munich (*6*), which enabled high-resolution lifetime measurements of phonons in various materials.

Bayrakci *et al.* present lifetime measurements of magnetic excitations obtained on the latest TAS–spin echo instrument built at the newly operational FRM-II reactor in Munich.

The author is at the Laboratory for Neutron Scattering, Eidgenössische Technische Hochschule (ETH Zurich) and the Paul Scherrer Institute, CH-5232 Villigen PSI, Switzerland. E-mail: joel.mesot@psi.ch

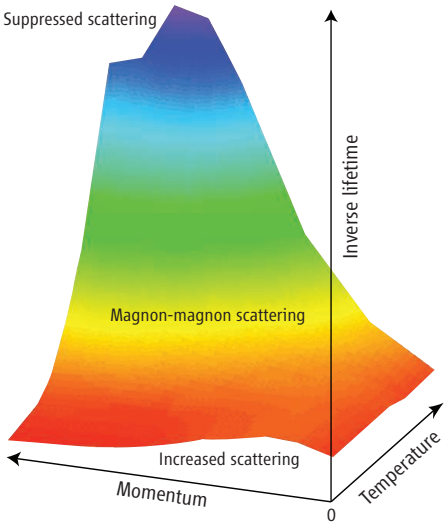




Published by AAAS

Even for the prototype antiferromagnetic $MnF_2$ compound studied by these authors, it is shown that the very high resolution ($\Delta E \approx 10^{-3}$ meV) that has been achieved allows one to unravel important discrepancies with well-established theories in the field of magnetism, as shown in the second figure. In this illustration, the vertical axis represents the measured energy width of a magnon excitation as a function of temperature and momentum. This quantity is inversely proportional to the magnon lifetime. For intermediate momenta and temperatures (green-yellow), the data can be explained by considering magnon-magnon scattering processes. On one hand, additional and unexpected scattering processes occur in the low-temperature regime (red), suggesting that new relaxation channels need to be considered. On the other hand, existing theoretical calculations of the magnon lifetime fail to reproduce the measured lifetime in the high-temperature and large-momentum regime (blue).

This work of Bayrakci *et al*. demonstrates that the TAS–spin echo technique proposed several years ago is now mature and ready for application to more complex materials like high-temperature superconductors or colossal-magnetoresistive manganites, where a delicate interplay exists among the magnetic, lattice, and electronic degrees of freedom. In a sense, neutron scattering is now seeing a revolution similar to the recent one in the field of ARPES with the development of high-resolution detectors ($\Delta E \approx$ 1 meV) and which has had a tremendous impact on our understanding of electron interactions in strongly correlated materials [for reviews, see (*7*, *8*)]. The unique capability to perform neutron scattering experiments with even higher resolutions over full dispersion relations opens completely new opportunities to test existing theoretical models of, and explore new phenomena in, solid state physics.




**Mapping magnetic interactions.** Three different scattering regimes of magnons in the antiferromagnet $MnF_2$ as determined by Bayrakci *et al*. using the TRISP spin-echo spectrometer at FRM-II.

Finally, it is worth mentioning that the TAS–spin echo device represents one of the many examples of recent technological advances realized at various neutron centers worldwide. This encompasses the development of new sources such as the third-generation spallation neutron sources in the United States, Japan, and hopefully soon in Europe, as well as the development of advanced neutron guide systems, polarization devices, new detectors, high-field facilities, and so on. These few examples illustrate that neutron scattering remains an active field and will continue to represent a unique and precious technique for the study of materials with novel electronic and magnetic properties.

**References and Notes**

1. L. D. Landau, *Sov. Phys. JETP* **3**, 920 (1957).
2. S. P. Bayrakci, T. Keller, K. Habicht, B. Keimer, *Science* **312**, 1926 (2006).
3. F. Mezei, *Z. Phys.* **255**, 146 (1972).
4. T. Keller, R. Golub, F. Mezei, R. Gähler, *Physica B* **234**, 1126 (1997).
5. M. Bleuel *et al., in Neutron Spin Echo Spectroscopy*, *Lecture Notes in Physics*, F. Mezei, C. Pappas, T. Gutberlet, Eds. (Springer-Verlag, New York, 2003), vol. 601.
6. T. Keller *et al., Appl. Phys. A* **74**, S332 (2002).
7. J. C. Campuzano, M. R. Norman, M. Randeria, in *Physics of Superconductors*, K. H. Bennemann, J. B. Ketterson, Eds. (Springer-Verlag, Berlin, 2004), vol. 2, pp. 167–273.
8. A. Damascelli, Z. X. Shen, Z. Hussain, *Rev. Mod. Phys*. **75**, 473 (2003).
9. Useful discussions with Ch. Mudry of the Condensed Matter Theory group at the Paul Scherrer Institute are acknowledged.

10.1126/science.1129459

ECOLOGY

# Climate Change and Crop Yields: Beyond Cassandra

An analysis of recent data from a wide variety of field experiments suggests that previous studies overestimated the positive effects of higher carbon dioxide concentrations on crop yields.

**David Schimel**

The effects of increasing atmospheric carbon dioxide ($CO_2$) concentrations on climate have been a source of worry for decades, but the positive effect of increasing $CO_2$ levels on crop growth has been a silver lining in the climate change cloud. Hundreds of studies, some dating back decades, have shown that most major food crops respond positively to increasing $CO_2$ concentrations, because of the direct stimulatory effect of $CO_2$ on photosynthesis and the indirect effect of decreasing the water requirements of crops. The former effect should make crops more productive and the latter more drought-tolerant. Yet a new analysis of recent and far more realistic studies based on the free-air concentration enrichment (FACE) technique, reported by Long *et al.* on page 1918 of this issue (*1*), sounds a cautionary note.

In FACE studies, a sophisticated, computer-controlled micrometeorological system enriches the air above a large plot to a target level, but does not otherwise alter the microclimate and growing environment. There is a large network of FACE studies around the world (see the figure), but relatively few have used food crops; most have been in natural or seminatural ecosystems. Long *et al.* synthesized the results from agronomic FACE studies and find the effects of increased $CO_2$ concentrations on crop yields to be no more than half of the levels expected from earlier laboratory and open-top chamber studies.

What is the significance of this result? In anticipating the impacts of global change on agriculture, crop models have been major tools. Modern crop models are widely used in agricultural research and planning and have been shown to be robust and accurate under a wide range of conditions in today's world. These models have been extended to study climate change scenarios, including simulated responses to climate and $CO_2$ and a wide range of assumptions about technological change and adaptation by farmers and markets.

These modeling results suggest that, at least in some regions, the future of agriculture may be one of the brighter spots of climate change. Although guarded in their remarks, most assessments of climate change impacts on agriculture have concluded that the global impacts are relatively small (although they may be severe in some areas) (*2*). This is largely because the assumed effects of $CO_2$ on

The author is in the Climate and Global Dynamics Division, National Center for Atmospheric Research, Boulder, CO 80307, USA. E-mail: schimel@ucar.edu

**The global network of FACE sites.** For a full list of sites, including contacts, see http://cdiac.ornl.gov/programs/FACE/whereisface.html.

crop growth and water use offset all or part of the negative impacts of warmer temperatures and rainfall changes (*3*). Depending on the region, this effect may ameliorate negative impacts of climate change or even allow for increases in crop yield.

Long *et al.* find that most current models are based on literature suggesting 20 to 30% increases in photosynthesis and yield. This literature also suggests substantial effects in maize. Maize, a C4 photosynthesis pathway crop, should have no direct response to $CO_2$ and should only respond via decreased water requirements (C4 photosynthesis uses a $CO_2$-concentrating mechanism at the cellular level that is less sensitive to atmospheric $CO_2$ than the more common C3 pathway).

The extant FACE results in food crops are different from earlier lab and chamber results in a consistent and disturbing way. In surveying the effects of $CO_2$ concentration in the more realistic FACE studies, Long *et al.* conclude that the actual effect of $CO_2$ concentration is about half of that in the experiments used to develop climate change models. Despite earlier reports of substantial effects in C4 crops, Long *et al.* found no increase in yield at all in this case, although there was some evidence for increased drought tolerance.

If Long *et al.* are right, then even the current guarded assessment of the impacts of climate change on food crops may be optimistic. If an assumption of a nearly 30% increase in yield at 550 parts per million by volume (ppmv) of $CO_2$ is buried in current models and if the true response is half that, then the global pattern of modeled agricultural yields could be an overly rosy picture. Areas with modeled increases might see no change, and areas with no change or reductions might in reality experience crop failure.

In 2001, the Intergovernmental Panel on Climate Change assessment concluded that "experiments have shown that relative enhancement of productivity caused by elevated $CO_2$ usually is greater when temperature rises but may be less for crop yields at above-optimal temperatures […] Although the beneficial effects of elevated $CO_2$ on the yield of crops are well established for the experimental conditions tested, this knowledge is incomplete for numerous tropical crop species and for crops grown under suboptimal conditions." [see p. 237 in (*2*)]. This conclusion may be overly optimistic, and concerns about food security, especially in regions with high temperatures, poor soils, and dry conditions, may need to be revisited.

The report by Long *et al.* may move impacts on agriculture higher up on the list of pressing concerns about climate change, but this is not where the authors conclude their analysis. Rather, they note that the theoretical effect of increased $CO_2$ on yield is much larger (36% at twice the pre-industrial $CO_2$ levels) than the realized effects on yield. The observed effect on yield is typically closer to 20%, with effects on biomass and yield being even smaller (20 and 13%, respectively).

This result calls for a cautious use of models, but also suggests an opportunity. Some set of biological processes appears to operate to reduce the impact of $CO_2$ on realized gains in biomass and yield below that expected from the effects on photosynthesis. The processes that cause this reduction are poorly understood and are not included effectively in models to this day. Long *et al.* conclude that this is an opportunity for crop breeders to develop varieties that can take advantage of the increase in atmospheric $CO_2$ concentration. Crop breeders, especially when able to use genetic engineering techniques, have produced an extraordinary array of crops for a multitude of uses and environments.

If successful, breeding for increased $CO_2$ concentration could become a major factor in the agricultural sector and could move agriculture back off the list of the most pressing concerns about climate change. Although Long *et al.* begin Cassandra-like by raising a new concern, they conclude with a creative suggestion for adaptation to increased $CO_2$ levels using the proven creativity of agronomic science.

**References**


1. S. P. Long, E. A. Ainsworth, A. D. B. Leakey, J. Nösberger, D. R. Ort, *Science* **312**, 1918 (2006).
2. J. J. MacCarthy *et al.*, Eds., *Climate Change 2001: Impacts, Adaptation and Vulnerability* (Cambridge Univ. Press, New York, 2001).
3. J. Reilly, D. Schimmelpfennig, *Clim. Change* **43**, 745 (1999).


10.1126/science.1129913

INTERNATIONAL

# Pioneering AAAS Project Uses Satellites to Aid Human Rights

On 14 May 2004, a satellite passing 450 km over Zimbabwe captured an image that included portions of the hardscrabble Hatcliffe settlement—more than 700 homes and other buildings scattered across the grasslands just north of the nation's capital city, Harare. Less than 16 months later, on 2 September 2005, a satellite sent back a stunning new picture: The pattern of red-dirt roads was still visible, but the buildings were gone.

The pictures were among the first collected by a commercial satellite company as part of a year-long AAAS pilot project to assess how satellites and other geospatial technology can be used in support of human rights. Already the project is having an impact: Amnesty International and Zimbabwe Lawyers for Human Rights used satellite pictures of the destroyed Porta Farm settlement in a 31 May report on the Zimbabwe government's destructive campaign to uproot opposition, generating extensive newspaper and broadcast coverage in Europe, Africa, and the United States.

Otto Saki, an attorney with the Zimbabwe lawyers group, said in e-mailed remarks that the satellite images may have "a phenomenal impact" in legal action over the systematic destruction of villages under the government of President Robert Mugabe.

"New satellite technology provides the unprecedented ability to document human rights abuses via a virtual 'eye in the sky'," said Larry Cox, executive director of Amnesty International USA. "With satellite projects like this one, we are gaining the ability to detect, publicize, and even prevent future human rights abuses from occurring in Zimbabwe and around the world."

Geospatial technology is not new—the development of hot air balloons and airplanes brought the use of aerial cameras; intelligence agencies have long used spy satellites; and scientists use such tools to study the weather and forest fires. But images from government satellites are not usually available in a timely way to human rights groups, and new images from privately owned satellites can cost $2000 or more.

Last December, the AAAS Science and Human Rights program obtained a $110,000 grant for a pilot project from the John D. and Catherine T. MacArthur Foundation.

Portions of the Hatcliffe settlement outside of Harare, Zimbabwe, on 14 May 2004.

The same part of Hatcliffe on 2 September 2005—with scores of buildings demolished.



Imaging satellites and other geospatial technology have been "vastly underutilized" in human rights work, said Lars Bromley, who has guided the project as a senior program associate in the AAAS Office of International Initiatives. "By handling all the technical and analytical aspects, AAAS allows groups like Amnesty and the lawyers to match their issue expertise with the power of the imagery. If we can smooth this relatively complicated process, the NGOs working to protect human rights around the world can see lots of benefits."

Among key partners in the effort are Amnesty International USA; the United Nations Special Advisor for the Prevention of Genocide; the Natural Resources Defense Council; the U.S. Holocaust Memorial Museum; the U.S. Campaign for Burma; and EQUITAS, the international center for human rights education. DigitalGlobe, a Colorado-based satellite image company, has been a partner and has provided images at a discounted price from its high-resolution QuickBird imaging satellites. Another satellite image company, GeoEye, also has provided generous support.

Bromley and others say that sophisticated commercial satellites and the increasing power of personal computers and the Internet have made the data more available than ever. The costs are likely to fall in coming years as more commercial imaging satellites are launched.

After AAAS finishes analyzing images from Hatcliffe, Porta Farm, and two other settlements, the project will turn to test cases in the Darfur area of Sudan and Burma. Published reports on the project's interest in allegations of wholesale destruction in Burma's Karen State have elicited a sharp rebuke in a newsletter controlled by the nation's government.

Over the years, AAAS's Science and Human Rights program has pioneered a number of initiatives to develop and promote the use of scientific methods to advance human rights, including forensic sciences, statistics, and social science methods. If the geospatial pilot project is successful, AAAS and its partners will explore how to make it permanent.

EDUCATION

# Senior Scientists and Engineers Bring Experience to Class

Don Rea had been a research chemist and a research director at NASA and the Jet Propulsion Laboratory. He had never dissected a frog, though—but there he was, closely watched by a troop of middle-schoolers, gamely grabbing a scalpel.

Don Rea

"I had also never looked at protozoa before, but I was able to set up the microscopes and help the kids observe these," Rea added proudly.

Last September, Rea was one of nine volunteers who entered seven middle schools and a high school in Montgomery County, Maryland, to help students with their science lessons, support teachers with a little extra information, and even build a few experiments. Rea organized the volunteer program as a project of Senior Scientists and Engineers (SSE), a AAAS-affiliated group of retired researchers dedicated to public service.

Rea and the SSE thought the schools might be a good place to fight back against science illiteracy and apathy. "The general populace is not well tuned in to science, and frequently takes positions that are contrary to the interest of good scientific development," Rea said.

"At a time when many political leaders—and many parents—recognize the need for a renewed national commitment to education in science, technology, and mathematics, this program is an inspiration," said Alan I. Leshner, AAAS chief executive officer. "If retired professionals were able to form partnerships with schools nationwide, this approach could help improve science literacy and give students further evidence that these are exciting, important fields."

After discussions with the Montgomery County schools, Senior Scientists and Engineers and AAAS signed an agreement with the schools for a 1-year pilot program. SSE then sent out a call for retired AAAS members who could help out in classes once a week.

Rea said the idea was to broaden science's appeal to the first generation of the 21st century. "We're aren't interested in providing tutoring to students who want to become scientists," he explained. "Our objective is to try to increase science understanding of the entire class."

Teachers and students alike welcomed the newcomers. "It has been wonderful to have a person from the community who cares about science and knows a lot about scientific concepts to share his expertise with the students," said Michelle Stanton, a Montgomery Village Middle School teacher.

Nina Hoffman, a teacher at Argyle Middle School, worked with volunteer Dave Weiss, a retired mechanical engineer. "His rapport with the students was wonderful," she said. "When we're doing labs, it's great for the kids—now they have two adults to ask questions. And he just didn't go and answer their questions. He asked questions in response to their questions to really make them think. They really do appreciate his knowledge and what he brings to the classroom."

Rea noticed that, too. "In one case, the teacher asked me to make an observation and I got a round of applause, which was totally unexpected," he joked.

Volunteers built a wave machine for one class and a metal ramp for a physics demonstration in Hoffman's class, but Rea said that the volunteers focus mostly on supporting the teacher's lessons rather than introducing new activities.

The program won rave reviews in its first year, and both teachers and volunteers said that a year of experience would make them even more effective. It will be expanded to more Montgomery County middle schools this fall. SSE also is sharing ideas with similar programs around the country, Rea said.

"There are over a million scientists and engineers over 60 around the country, and the number is probably growing every year," Rea said. "That's a big pool we can draw on to improve science teaching in schools."

For more information on Senior Scientists and Engineers, visit www.seniorscientist.org

—*Becky Ham*

AAAS

## AAAS Annual Election

The slate of candidates for the 2006 election of AAAS officers will be announced in News and Notes in the 28 July issue of *Science*.

COMMUNICATION

## Screeners Needed for AAAS Science Journalism Awards

Scientist volunteers are needed to review entries in the prestigious AAAS Science Journalism Awards program. Scientists residing in the Washington, D.C., area, or who will be in the area in mid-August to mid-September, are invited to help screen print, radio, and television reports for scientific accuracy. If interested, please contact Lonnie Shekhtman (202-326-6434; lshekhtm@aaas.org) in the AAAS Office of Public Programs.

Winners of the awards, which are sponsored by Johnson & Johnson Pharmaceutical Research & Development, L.L.C., will be honored at a ceremony in February 2007 at the AAAS Annual Meeting in San Francisco. Members of the screening committees will be recognized in the awards booklet distributed during the ceremony.

AAAS

## New AAAS Dues Rates Approved for 2007

The AAAS Board of Directors has approved a dues increase for 2007. The Board authorizes increases to cover two kinds of expenses: unavoidable costs associated with running AAAS and publishing *Science*, and new expenses that add value to membership. Postage and paper increases and improving online resources are examples of the kind of expenses that the Board anticipated in setting the 2007 rates.

The new rates are effective for membership terms beginning after 31 December 2006. As listed below, they do not include postage or taxes for international members, which is additional.

| | |
|---|---|
| • Regular professional members | $142 |
| • Postdocs and K-12 teachers | $99 |
| • Emeritus members who receive print *Science* | $110 |
| • Students | $75 |
| • Patrons | $310 |
| • Supporting and Emeritus members who do not receive *Science* | $56* |

The Board also set the institutional subscription rate for print *Science* at $360 for high school and public libraries and $710 for all other institutions. For further information, including subscription rates for *Science* Online, librarians should contact AAAS or their subscription agents, or go to www.sciencemag.org/subscriptions/inst_sub.dtl on the Web.

All members will be advised of the new dues rates on their renewal notices for 2007. Member dues and voluntary contributions form the critical financial base for a wide range of AAAS activities. For more information, contact the AAAS Membership Office at 202-326-6417, www.aaas.org/membership/.

* Supporting member dues rate is set by the membership department

SCIENCE VALUES

## NAS, AAAS Launch Integrity Web Site

In a move to provide practical resources on ethics and integrity, AAAS and the National Academy of Sciences have joined to offer an online bibliography on scientific integrity.

The collaboration will bring an array of articles, essays, and other materials published by NAS and AAAS to young scholars, working scientists, and educators. The bibliography is based on the idea that trust and accountability are integral to the research enterprise and the sharing of scientific information.

Visit the bibliography at www.aaas.org/integrity.

# Bacterial Diversity in Tree Canopies of the Atlantic Forest

M. R. Lambais,[1]* D. E. Crowley,[3]* J. C. Cury,[1] R. C. Büll,[1] R. R. Rodrigues[2]

The leaf surface, also known as the phyllosphere, is one of the most common habitats for terrestrial microorganisms (*1*), but almost nothing is known about the diversity of microorganisms that inhabit this environment (*2*). Here, we report a survey of bacterial diversity in the leaf canopy of a tropical Atlantic forest. The Atlantic Forest of Brazil is a biodiversity hotspot that has been reduced to less than 8% of its original size over the past 4 centuries and is considered to be the oldest forest on the planet, containing about 20,000 vascular plant species, of which about one-half are endemic (*3*, *4*). Initially we compared the bacterial communities on the leaf surfaces of nine tree species (table S1) by using a molecular method that generates a DNA fingerprint of the predominant bacteria from their 16S ribosomal RNA (rRNA) gene sequences (*5*) (fig. S1). Our results showed that bacterial communities from the same tree species varied but could be consistently grouped by discriminant analyses (table S2). These data are consistent with previous research showing that different plants select for distinct microbial communities (*6*).

To identify the bacteria in the phyllospheres of *Trichilia catigua*, *T. clausenii*, and *Campomanesia xanthocarpa*, we analyzed 418 partial DNA sequences encoding 16S rRNA genes (*5*). Comparison of homologous and heterologous coverage curves indicated that all three phyllosphere communities were significantly different in their bacterial species compositions (fig. S2). For all clone libraries, the sample size was sufficient to recover the most abundant deep phylogenetic groups. At evolutionary distances (*D*) higher than 0.20 (the cutoff value for group sequences at the phylum level), the homologous coverages were greater than 96% (fig. S2). At a *D* of 0.03, corresponding to bacterial species, coverages varied from 67 to 81% (fig S2). Species richness was estimated by using Chao1 nonparametric estimator (table S3). Each phyllosphere community harbored from at least 95 to 671 bacterial species (Fig. 1 and table S3), of which only 0.5% were common to all of the trees studied. Almost all of the bacteria (97%) were from undescribed species, suggesting they may be unique to the phyllosphere habitat (table S4).

Although this initial survey was limited in scope, extrapolation of our results for the 20,000 vascular plant species in the Atlantic Forest would yield about 2 to 13 million new bacterial species. The absolute diversity of bacteria in nature is unknown, but by comparison the Earth's oceans have been estimated to contain up to 2 million species, whereas a ton of soil may have 4 million species (*7*). The estimates for phyllosphere diversity could be decreased considerably should future surveys reveal higher amounts of overlap in bacterial community composition between tree species. On the other hand, the bacterial species richness for the individual trees surveyed represent minimum estimates of that which may occur on individual trees. Variations in community structures within tree species may possibly correspond to different leaf ages, location in the canopy, light incidence, and microclimate conditions that influence the leaf environment (*8*). The current study provides a glimpse into the microbial diversity in tree canopies of tropical forests, and there are many questions that arise from this research. Do the same tree species in completely different locations or continents harbor similar communities? To what degree do various environmental factors affect the composition and structure of phyllosphere communities? What is the diversity of fungi and archaea on the plant leaf surfaces, and what role does phyllosphere microbial community play in protection against herbivory or infection by pathogens? As we begin to survey the bacterial species through systematic surveys of different plants, there will be exciting opportunities for studies of the metabolic capabilities and the ecological functions of phyllosphere microorganisms in terrestrial ecosystems.

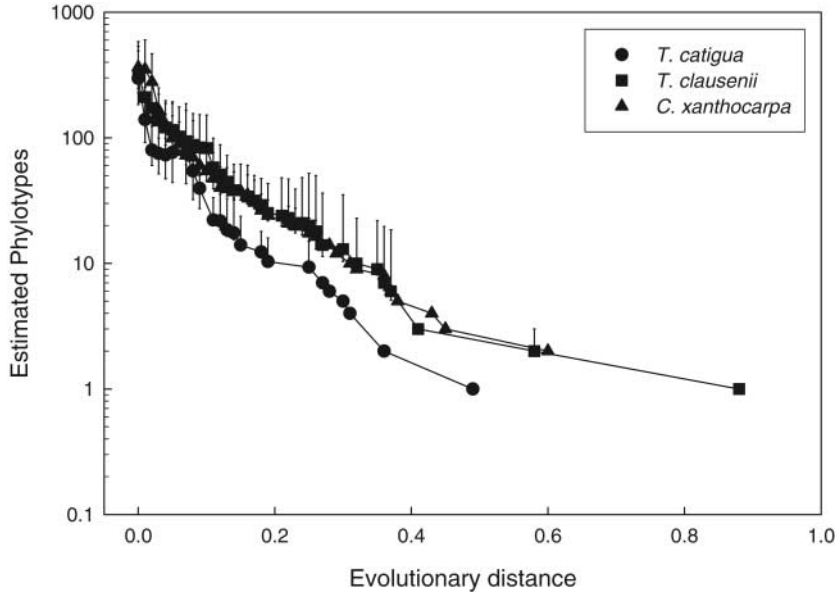



**Fig. 1.** Estimated number of phylotypes at different evolutionary distances, using Chao1 nonparametric estimator, on the leaf surface of different tree species. Bars represent 95% confidence intervals.

### References and Notes

1. J. Ruinen, *Nature* **177**, 220 (1956).
2. S. E. Lindow, M. T. Brandl, *Appl. Environ. Microbiol.* **69**, 1875 (2003).
3. Ecosystem Profile for Atlantic Forest Biodiversity Hotspot, Conservation International Critical Ecosystem Partnership Fund, 2001 (www.cepf.net/xp/cepf/static/pdfs/Final.AtlanticForest.EP.pdf).
4. N. Myers, R. A. Mittermeier, C. G. Mittermeier, G. A. B. da Fonseca, J. Kents, *Nature* **403**, 853 (2000).
5. Materials and methods are available on *Science* Online.
6. C. H. Yang, D. E. Crowley, J. Borneman, N. T. Keen, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 3889 (2001).
7. T. P. Curtis, W. T. Sloan, J. W. Scannell, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 10494 (2002).
8. R. K. P. Yadav, K. Karamanoli, D. Vokou, *Microb. Ecol.* **50**, 185 (2005).
9. We acknowledge C.-H. Yang and G. Sparovek for contribution and discussion of ideas, G. Franco for assistance in identification of the tree species, and M. Giannoti and N. Ivanauskas for logistical support. This project was supported by grant 99/09635-0 from Fundação de Amparo à Pesquisa do Estado de São Paulo, as part of the BIOTA-FAPESP, the Biodiversity Virtual Institute Program (www.biota.org.br). All nucleotide sequences have been deposited at GenBank under the accession numbers DQ221265 to DQ221691.

### Supporting Online Material

www.sciencemag.org/cgi/content/full/312/5782/1917/DC1
Materials and Methods
Figs. S1 and S2
Tables S1 to S4
Data and Analyses

6 January 2006; accepted 8 May 2006
10.1126/science.1124696

[1]Department of Soils and Soil Science, [2]Department of Biological Sciences, University of São Paulo, Piracicaba, São Paulo, Brazil. [3]Department of Environmental Sciences, University of California, Riverside, CA 92591, USA.

*To whom correspondence should be addressed. E-mail: mlambais@esalq.usp.br (M.R.L.); crowley@ucr.edu (D.E.C.)

# Food for Thought: Lower-Than-Expected Crop Yield Stimulation with Rising $CO_2$ Concentrations

Stephen P. Long,[1,2,3]* Elizabeth A. Ainsworth,[4,1,3] Andrew D. B. Leakey,[3,1] Josef Nösberger,[5] Donald R. Ort[4,1,2,3]

Model projections suggest that although increased temperature and decreased soil moisture will act to reduce global crop yields by 2050, the direct fertilization effect of rising carbon dioxide concentration ($[CO_2]$) will offset these losses. The $CO_2$ fertilization factors used in models to project future yields were derived from enclosure studies conducted approximately 20 years ago. Free-air concentration enrichment (FACE) technology has now facilitated large-scale trials of the major grain crops at elevated $[CO_2]$ under fully open-air field conditions. In those trials, elevated $[CO_2]$ enhanced yield by ~50% less than in enclosure studies. This casts serious doubt on projections that rising $[CO_2]$ will fully offset losses due to climate change.

Much effort has been put into linking models of climate and crop growth to project future changes in crop yields and food supply across the globe (*1*–*4*). Projections reviewed by the Intergovernmental Panel on Climate Change (IPCC) suggest that increased temperature and decreased soil moisture, which would otherwise reduce crop yields, will be offset by the direct fertilization effect of rising carbon dioxide concentration ($[CO_2]$) (*5*–*7*). The IPCC projections suggest that total crop yield may rise when averaged across the globe, but this net gain will result from generally lower yields in the tropics and increased yields in temperate zones. The accuracy of these projections and thus future food security depend critically on the magnitude of the $CO_2$ fertilization effect under actual growing conditions.

Atmospheric $[CO_2]$ has risen from ~260 parts per million (ppm) approximately 150 years ago to 380 ppm today (*8*). Yet $[CO_2]$ is markedly uniform across the globe; so, in contrast to temperature and soil moisture, there is no consistent spatial variation on which to estimate yield responses to increasing $[CO_2]$. Similarly, it is not easy to alter $[CO_2]$ experimentally around a crop in the field. As a result, most information about crop responses to elevated $[CO_2]$ is obtained from studies in greenhouses, laboratory controlled-environment chambers, and transparent field chambers, where released $CO_2$ may be retained and easily controlled. These settings have provided the basis for projecting $CO_2$ fertilization effects on the major food crops: maize, rice, sorghum, soybeans, and wheat.

Crops sense and respond directly to rising $[CO_2]$ through photosynthesis and stomatal conductance, and this is the basis for the fertilization effect on yield (*9*). In $C_3$ plants, mesophyll cells containing ribulose-1,5-bisphosphate carboxylase-oxygenase (RuBisCO) are in direct contact with the intercellular air space that is connected to the atmosphere via stomatal pores in the epidermis. Hence, in $C_3$ crops, rising $CO_2$ increases net photosynthetic $CO_2$ uptake because RuBisCO is not $CO_2$-saturated in today's atmosphere and because $CO_2$ inhibits the competing oxygenation reaction leading to photorespiration. RuBisCO is highly conserved across terrestrial plants, so instantaneous responses to increased $[CO_2]$ may be generalized across $C_3$ plants, including rice, soybeans, and wheat. In theory, at 25°C, an increase in $[CO_2]$ from the present-day value of 380 ppm to that of 550 ppm, projected for the year 2050, would increase $C_3$ photosynthesis by 38% (*9*). In contrast, in $C_4$ crops such as maize and sorghum, RuBisCO is localized to bundle sheath cells in which $CO_2$ is concentrated to three to six times atmospheric $[CO_2]$ (*10*). This concentration is sufficient to saturate RuBisCO and in theory would prevent any increase in $CO_2$ uptake with rising $[CO_2]$. Although $C_4$ crops may not show a direct response in photosynthetic activity, an indirect increase in the efficiency of water use via reduction in stomatal conductance may still increase yield (*9*).

**How have $CO_2$ fertilization factors been derived?** Most models used to predict future crop yields, including those within the IPCC (*5*), are from two families: the Decision Support System for Agrotechnology Transfer (DSSAT) (*6*, *11*, *12*) and the Erosion Productivity Impact Calculator (EPIC) (*13*–*15*). Studies using DSSAT assume $CO_2$ fertilization factors based on the method of Peart *et al.* (*3*), which used summaries for soybeans (*16*), maize (*17*), wheat (*18*), and rice (*18*). Studies using EPIC (*13*–*15*) assume $CO_2$ fertilization factors based on the method of Stockle *et al.* (*4*), which parameterized a $CO_2$ response function to reproduce the mean yield stimulations reported for elevated $[CO_2]$ by Kimball (*18*). Tracing DSSAT and EPIC methods back reveals that the magnitude of the $CO_2$ fertilization effects in these models is primarily based on data from three literature reviews from the 1980s (*16*–*18*). The $CO_2$ fertilization effects reported in these reviews for the major crops are given in Table 1

**Table 1.** Percentage increases in yield, biomass, and photosynthesis of crops grown at elevated $[CO_2]$ (550 μmol $mol^{-1}$) relative to ambient $[CO_2]$ in enclosure studies versus FACE experiments. Data for enclosure studies were summarized by Kimball (*18*), Cure and Acock (*17*), and Allen *et al.* (*16*) and in Fig. 2. Mean response ratios from these reviews were adjusted to an elevated $[CO_2]$ of 550 μmol $mol^{-1}$ by means of the nonrectangular hyperbolic functions for $C_3$ and $C_4$ species from Fig. 2. The values that summarize all chamber studies shown in Fig. 2 are given in the row entitled "enclosure studies." Percentage increases for FACE studies were generated by meta-analysis [see supporting online material (SOM) and table S2] (*37*).

| Source | Rice | Wheat | Soybeans | $C_4$ crops |
|---|---|---|---|---|
| | | *Yield* | | |
| Kimball (1983) | 19 | 28 | 21 | – |
| Cure and Acock (1986) | 11 | 19 | 22 | 27 |
| Allen *et al.* (1987) | – | – | 26 | – |
| Enclosure studies | – | 31 | 32 | 18 |
| FACE studies | 12 | 13 | 14 | 0* |
| | | *Biomass* | | |
| Cure and Acock (1986) | 21 | 24 | 30 | 8 |
| Allen *et al.* (1987) | – | – | 35 | – |
| FACE studies | 13 | 10 | 25 | 0* |
| | | *Photosynthesis* | | |
| Cure and Acock (1986) | 35 | 21 | 32 | 4 |
| FACE studies | 9 | 13 | 19 | 6 |

*Data from only 1 year in Leakey *et al.* (*30*).

[1]Department of Plant Biology, [2]Department of Crop Sciences, [3]Institute for Genomic Biology, University of Illinois at Urbana Champaign, 1201 West Gregory Drive, Urbana, IL 61801, USA. [4]Photosynthesis Research Unit, U.S. Department of Agriculture–Agricultural Research Service, 1201 West Gregory Drive, Urbana, IL 61801, USA. [5]Institute for Plant Sciences, ETH Zurich, 8902 Zurich, Switzerland.

*To whom correspondence should be addressed. E-mail: stevel@life.uiuc.edu

after adjustment to estimate crop performance at a common [$CO_2$] of 550 ppm. Collectively, the fertilization factors averaged across the $C_3$ crops (rice, wheat, and soybeans) are 24% for yield, 27% for biomass, and 29% for photosynthesis. The responses for maize were lower except for yield, which was reported to increase by 27% (Table 1). All studies included in the reviews used enclosures, such as controlled environmental chambers, transparent field enclosures, or open-top chambers. Since the 1980s, many further chamber studies have been conducted. When these are compiled for wheat and soybeans, an even larger yield fertilization factor of 31% is suggested (Table 1). Although this is a wealth of data on which to project a $CO_2$ fertilization effect for crops across the globe, no agrochemical or plant-breeding company would base its business plan for a new chemical or variety solely on greenhouse studies without rigorous field trials (*19*, *20*). Yet our current projections of future world food supply are based on such potentially inadequate data.

**Why might chamber studies be inadequate for predicting future yields?** Many chamber studies used plants grown in pots, which are now known to alter the response of plants to elevated [$CO_2$] (*21*). Most of the field studies used open-topped and transparent-walled chambers, up to 2 m in diameter. Despite being partially open to the atmosphere, important environmental differences remain. In a chamber carefully designed to minimize environmental differences, receiving ~75% of full sunlight, the temperature inside the chamber was 4.3°C warmer and the water vapor pressure deficit was 0.8 kPa higher (*22*) than outside the chamber. The transmission of sunlight into the chambers was lower and the ratio of diffuse to direct sunlight increased. Other chamber types would cause even greater perturbation of the natural environment. All chambers alter air flow and intercept rainfall. Access by pests and diseases is restricted, but if they gain access, higher humidity and more shelter may accentuate epidemics. As a result, the effect of the chamber on plants is often greater than that of elevated [$CO_2$] (*23*). In agronomic trials, buffer rows are used between treatments; typically the width of this zone is twice the height of the crop. Because of the small practical size of chambers, most or all of the treated crop will be within this zone, which could exaggerate the response to elevated [$CO_2$] (*23*). To overcome these limitations, free-air concentration enrichment (FACE) was developed.

**How does FACE work?** A typical FACE apparatus consists of a 20-m-diameter plot within the crop field (Fig. 1A), in which $CO_2$ is released just above the crop surface on the upwind side of the plot. Wind direction, wind velocity, and [$CO_2$] (or ozone concentration) are measured at the center of the plot. Fast-feedback computer control then adjusts the positions and amount of $CO_2$ released at different points around the plot. These systems have been engineered so that they can operate continuously from sowing to harvest and maintain [$CO_2$] within the plot to within ±10% of the target level, either 550 or 600 ppm, for ~90% of the time (*9*, *24*–*26*) (Fig. 1B). Elevated [$CO_2$] decreases transpiration and therefore evaporative cooling, so that in sunlight the crop is warmer. This can serve to illustrate the uniformity of treatment (Fig. 1B).

Mini-FACE systems as small as 1 m in diameter have been developed and have proved invaluable in ecosystem studies where the focus is on the effect of increased input of carbon (*27*), but they do not escape the problems of enclosures with respect to scale. Avoiding edge effects associated with small plots is critical when the objective is to determine an exact $CO_2$ fertilization factor for crops. Our analysis has therefore been limited to full-size FACE systems of plots >8 m in diameter, investigating the five major global food crops and managed pasture systems (table S1): wheat at Maricopa, Arizona, USA, in 1992, 1993, 1996, and 1997; managed grassland at Eschikon, Switzerland, from 1993 to 2002; managed pasture at Bulls, New Zealand, from 1997 to 2002; sorghum at Maricopa, Arizona, USA, in 1998 and 1999; rice at Shizukuishi, Japan, from 1998 to 2000; and soybeans at Urbana-Champaign, Illinois, USA, from 2001 to 2005 and maize at the same location in 2002 and 2004 (*26*, *28*).

**What have we learned from the FACE experiments?** The response of plant production to [$CO_2$] is approximately hyperbolic, increasing linearly at subambient concentration and saturating at around 800 to 2000 ppm. The ratio of yield at treatment [$CO_2$] to yield at atmospheric [$CO_2$] was calculated for over 340 independent chamber studies. Hyperbolas of the response of yield to [$CO_2$] were then fit for wheat, soybeans, and $C_4$ grains (maize and sorghum combined) (Fig. 2). Only one replicated FACE experiment was conducted with each of these crops, but these experiments were repeated over 2 to 5 years. It was notable that for each crop, the stimulation of yield observed in FACE experiments fell well below (about half) the value predicted from chambers (Fig. 2). This was apparent for total biomass and most marked for photosynthesis. Notably, the stimulation of photosynthesis by elevated [$CO_2$] in enclosure studies of rice was four times the value observed in the rice FACE experiment (Table 1). With so few FACE studies, it might be thought that these lower values are the result of chance. Table 1 shows that for three key production measures in four crops, only 1 of the 12 items is not lower than the chamber equivalent. The probability of this outcome being attributed to chance is remote ($P = 0.003$).

Results from FACE experiments with $C_4$ crops are consistent with $CO_2$ having no direct effect on photosynthesis, but there may be an indirect effect through the amelioration of drought stress by reduced stomatal conductance at elevated [$CO_2$] (*29*–*31*). This fits the theoret-

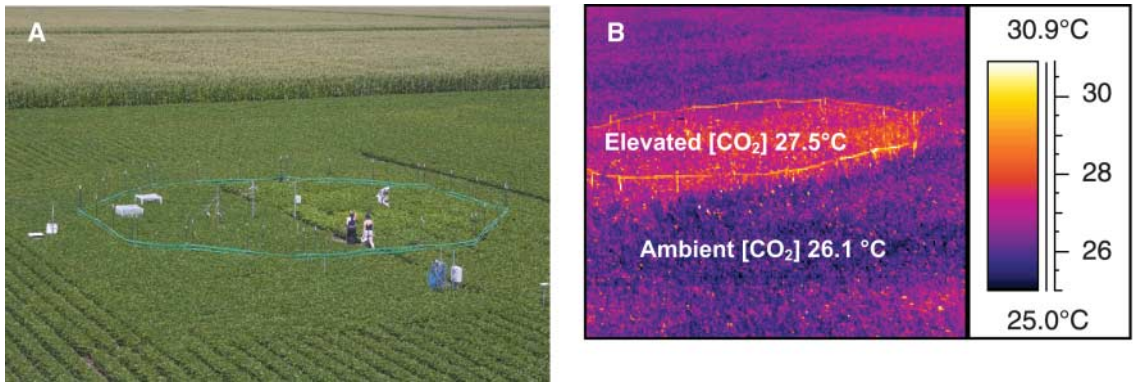



**Fig. 1.** (**A**) One of the 16 FACE plots of soybeans at the University of Illinois SoyFACE facility. $CO_2$ is released into the wind from nozzles in the green pipe, on the upwind side of the plot. Release rate is determined by wind speed and [$CO_2$], which is measured at the center of each ring. (**B**) The false-color infrared image provides a simple visualization of the uniformity of $CO_2$ treatment within a FACE plot. Here the atmosphere around a maize crop within the octagonal plot is maintained at 550 ppm [$CO_2$], whereas the rest of the field is at the current ambient [$CO_2$] of ~380 ppm. Maize growing inside an elevated [$CO_2$] plot was warmer in full sunlight than maize growing under ambient [$CO_2$] outside the plot at 15:30 on 15 July 2004. At that time, the average canopy temperature inside the four elevated [$CO_2$] plots at SoyFACE was 27.9 ± 0.2°C, significantly higher than canopy temperatures under ambient [$CO_2$] outside the plots (26.8 ± 0.3°C; $P = 0.03$). Because the pipes surrounding the plot are dry, they are warmer and so appear as white or light yellow. Greater canopy temperatures under elevated [$CO_2$] result from lower stomatal conductance, reducing latent heat loss by evapotranspiration and leading to lower crop water use, as described in Leakey *et al.* (*30*).

ical expectation that $C_4$ photosynthesis is $CO_2$-saturated at current atmospheric [$CO_2$] (*10*); therefore, no yield increase would be expected for well-watered crops. Under drought, elevated [$CO_2$] increased midday photosynthesis by 23% in sorghum (*31*). This failed to translate into a significant yield increase (*32*). On average, no significant yield increase has been observed for $C_4$ crops or $C_4$ wild grasses at elevated [$CO_2$] in FACE studies (*28*). This is in sharp contrast to the large stimulation of yield for well-watered plants in chambers (Fig. 2B) used to parameterize models. This suggests that the consistent stimulation of $C_4$ crop yield by elevated [$CO_2$] currently applied in models is inappropriate. At best, yield will in all probability be enhanced by elevated [$CO_2$] only in times and places of drought.

Wheat and rice FACE experiments included nitrogen treatments. At the lowest [N] (15 to 70 kg of N $ha^{-1}$), the average yield increase with elevated [$CO_2$] was only 9% (*28*), just over one-third of that of the chamber response (Table 1). Although this N input treatment was considered low by the standards of intensive agriculture in the European Union and United States, these levels exceed the world average and may therefore be closer to the stimulation factor for crop yields across the globe. Lower-than-expected yields under elevated [$CO_2$] are not just confined to grain crops. For example, the major $C_3$ herbage grass, *Lolium perenne*, also showed a yield increase of only 9% at two locations; and at the lowest [N] (100 to 140 kg of N $ha^{-1}$), the yield increase was an insignificant 1% (table S2) (*28*). Although the data here apply to a single species, *L. perenne* is one of the most important and widely grown herbage grasses in the temperate zone.

No FACE experiment has been conducted in the tropics, but two factors emerging from temperate studies have particular implications for tropical crops. First, the $CO_2$ fertilization effect may be small without large additions of N. Second, FACE experiments with the major grain crops of sub-Saharan Africa, sorghum and maize, have so far failed to show any yield increase from elevated [$CO_2$]. Parry *et al*. (*7*) projected that yield losses in these countries due to climate change could be 10 to 30% by 2050, but these would be ameliorated to only 2.5 to 5% when the $CO_2$ fertilization effect is added (*7*). The FACE experiments suggest that this amelioration may be far less than expected.

**Rising surface ozone.** Increased combustion of fuels will increase not only atmospheric [$CO_2$] but also atmospheric nitrogen oxide concentrations, which, when coupled with climate change, will result in a continued increase in surface ozone concentration ([$O_3$]). Many rural areas in the temperate zone of the Northern Hemisphere, as well as in the tropics, are forecast to see increases in [$O_3$] of ~20% by midcentury (*8*). Ozone is toxic to plants at concentrations as low as 30 parts per billion (ppb). Although chamber studies have shown large yield losses owing to elevated [$O_3$] (*33*), these effects are not incorporated in current projections of future yields (*2*, *8*).

Until very recently, the only studies of the effects of elevated [$O_3$] on crops were conducted in chambers, and it was unclear whether similar losses would occur under conditions of normal canopy/atmosphere coupling in the field. Morgan *et al*. (*34*) used a FACE system adapted to elevate [$O_3$] rather than [$CO_2$] to examine whether the decreases in yield for soybeans in central Illinois projected from chamber experiments occurred in the open air. A 23% increase in [$O_3$] from an average daytime ambient concentration of 56 to 69 ppb over two growing seasons decreased soybean yield by 20%. How does this compare with the expectations established from chamber studies? Based on a prior compilation of chamber studies (*33*), the expected decrease was 8%. If the effects of [$CO_2$] and [$O_3$] observed in FACE studies are additive, then the net effect of simultaneous increases in [$O_3$] and [$CO_2$], as forecast by the IPCC A1B scenarios, would be a 5% decrease in yield, compared with the 23% increase used to parameterize current models (Table 1). Chamber studies suggest that elevated [$CO_2$] may provide some protection against elevated [$O_3$] and therefore the effects will not be additive, but this has yet to be verified for any crop under open-air field conditions.

**What is needed?** The $CO_2$ fertilization effects, derived from chamber experiments, currently used in crop models forecast substantial increases in future crop production under conditions associated with climate change. The FACE experiments, conducted in open fields, are not without their limitations (*26*, *35*), but represent our best simulations of the future elevated [$CO_2$] environment. Our meta-analytic summary of the FACE experiments indicates that there will be a much smaller $CO_2$ fertilization effect on yield than currently assumed, and possibly little or no stimulation for $C_4$ crops.

The average yield increase at elevated [$CO_2$] for crops in FACE studies fell well short of the

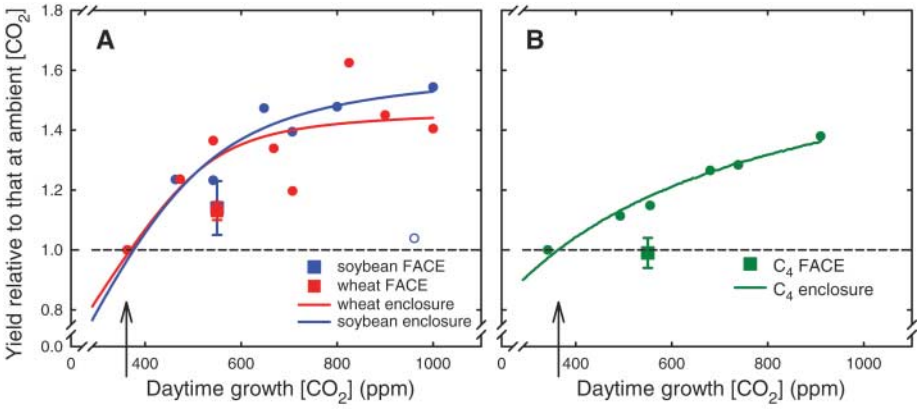



**Fig. 2.** Effects of elevated [$CO_2$] on crop yield. Data are yields at elevated [$CO_2$] relative to those at ambient [$CO_2$] (arrow) for (**A**) soybeans in chambers (solid blue circles) and FACE (blue square, hidden behind red square) and wheat in chambers (red circles) and FACE (red square); and (**B**) $C_4$ crops (maize and sorghum combined) in chambers (green circles) and FACE studies (green square). Error bars indicate mean ± 90% confidence intervals around the means for the FACE studies. The chamber studies included 115 independent measures of soybeans (*21*), 211 of wheat (*36*), and 14 of maize and sorghum (table S3). These measures were divided into 10 classes of growth [$CO_2$] in 100-ppm increments. Plotted values are the class means of growth [$CO_2$] and yield. Solid lines are the least-squares fits for the nonrectangular hyperbolic response of yield to growth [$CO_2$] from these enclosure studies of soybeans (blue line, $r^2 = 0.98$), wheat (red line, $r^2 = 0.88$), and $C_4$ crops (green line, $r^2 = 0.99$). The yield response of soybeans in chambers to growth [$CO_2$] of 900 to 999 ppm [open blue circle in (A)] was an outlier and was excluded from the curve fitting. Full details of the meta-analysis methods and results from FACE are presented in the SOM and table S2.

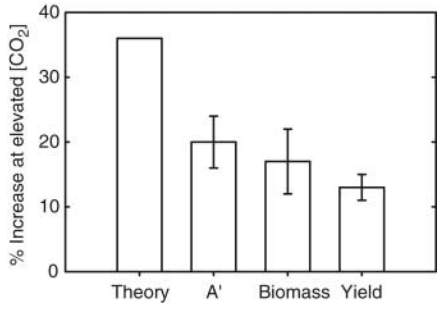



**Fig. 3.** Comparison of theoretical and actual changes in $C_3$ crop production parameters at an elevated [$CO_2$] of 550 ppm relative to ambient [$CO_2$]. Theory, theoretical RuBisCO-limited photosynthesis at 550 ppm [(*9*) and SOM]; A′, measured daily integral of carbon uptake; biomass, final above-ground biomass; yield, harvestable grain yield. Error bars indicate mean ± 90% confidence intervals. A′, biomass, and yield were measured in $C_3$ crops exposed to elevated [$CO_2$] in FACE experiments (table S2).

theoretically possible increase based on the well-defined properties of RuBisCO (Fig. 3). At 25°C, an increase in [$CO_2$] to 550 ppm should increase light-saturated photosynthesis by 36%. The average increase observed for $C_3$ crops in FACE was 20% for the daily integral of photosynthetic $CO_2$ uptake, 17% for total biomass, and just 13% for yield (Fig. 3). This suggests that a series of feedbacks operate in the field to constrain realization of the potential benefits of elevated [$CO_2$]. Only with a thorough high-priority R&D effort might we overcome these feedbacks and achieve the potential gains in food supply.

The FACE experiments clearly show that much lower $CO_2$ fertilization factors should be used in model projections of future yields; however, the present experiments are limited in the range of growing conditions that they cover. Scientists have not investigated the interactive effects of simultaneous change in [$CO_2$], [$O_3$], temperature, and soil moisture. Technological advances suggest that large-scale open-air facilities to investigate these interactions over controlled gradients of variation are now possible (*26*). Although we have projected results to 2050, this may be too far in the future to spur commercial R&D, but it must not be seen as too distant to discourage R&D in the public sector, given the long lead times that may be needed to avoid global food shortage.

### References and Notes


1. G. Hoogenboom *et al.*, in *Climate Change and Agriculture: Analysis of Potential International Impacts. ASA Special Publication no. 59* (American Society of Agronomy, Madison, WI, 1995), pp. 51–75.
2. M. Parry, C. Rosenzweig, M. Livermore, *Philos. Trans. R. Soc. London Ser. B* **360**, 2125 (2005).
3. R. M. Peart, J. W. Jones, R. B. Curry, K. J. Boote, L. H. Allen, in *The Potential Effects of Global Climate Change on the United States, Appendix C, Report to Congress*, J. B. Smith, D. A. Tirpak, Eds. (EPA-230-05-89-053, U.S. Environmental Protection Agency, Washington, DC, 1989), pp. 2–54.
4. C. O. Stockle, J. R. Williams, N. J. Rosenberg, C. A. Jones, *Agric. Syst.* **38**, 225 (1992).
5. H. Gitay, S. Brown, W. Easterling, B. Jallow, in *Climate Change 2001: Impacts, Adaptation, and Vulnerability*, J. J. McCarthy, O. F. Canziani, N. A. Leary, D. J. Dokken, K. S. White, Eds. (Cambridge Univ. Press, Cambridge, 2001), pp. 237–342.
6. M. Parry, C. Rosenzweig, A. Iglesias, G. Fischer, M. Livermore, *Global Environ. Change* **9**, S51 (1999).
7. M. L. Parry, C. Rosenzweig, A. Iglesias, M. Livermore, G. Fischer, *Global Environ. Change* **14**, 53 (2004).
8. J. T. Houghton *et al.*, Eds., *Climate Change 2001: The Scientific Basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change* (Cambridge Univ. Press, Cambridge, 2001).
9. S. P. Long, E. A. Ainsworth, A. Rogers, D. R. Ort, *Annu. Rev. Plant Biol.* **55**, 591 (2004).
10. S. von Caemmerer, R. T. Furbank, *Photosynth. Res.* **77**, 191 (2003).
11. R. M. Adams *et al.*, *Nature* **345**, 219 (1990).
12. C. Rosenzweig, A. Iglesias, in *Understanding Options for Agricultural Production*, G. Y. Tsuji, G. Hoogenboom, P. K. Thornton, Eds. (Kluwer Academic, Dordrecht, Netherlands, 1998), pp. 267–292.
13. R. A. Brown, N. J. Rosenberg, *Clim. Change* **41**, 73 (1999).
14. R. C. Izaurralde, N. J. Rosenberg, R. A. Brown, A. M. Thomson, *Agric. For. Meteorol.* **117**, 97 (2003).
15. A. M. Thomson, R. A. Brown, N. J. Rosenberg, R. C. Izaurralde, V. Benson, *Clim. Change* **69**, 43 (2005).
16. L. H. Allen *et al.*, *Global Biogeochem. Cycles* **1**, 1 (1987).
17. J. D. Cure, B. Acock, *Agric. For. Meteorol.* **38**, 127 (1986).
18. B. A. Kimball, *Agron. J.* **75**, 779 (1983).
19. A. Anand *et al.*, *J. Exp. Bot.* **54**, 1101 (2003).
20. B. Black, *Abstr. Pap. Am. Chem.* **228**, U84 (2004).
21. E. A. Ainsworth *et al.*, *Global Change Biol.* **8**, 695 (2002).
22. D. Whitehead *et al.*, *J. Biogeogr.* **22**, 307 (1995).
23. A. R. McLeod, S. P. Long, *Adv. Ecol. Res.* **28**, 1 (1999).
24. F. Miglietta, M. Lanini, M. Bindi, V. Magliulo, *Global Change Biol.* **3**, 417 (1997).
25. F. W. Lipfert, Y. Alexander, G. R. Hendrey, K. F. Lewin, J. Nagy, *Crit. Rev. Plant Sci.* **11**, 143 (1992).
26. J. Nösberger *et al.*, Eds. *Managed Ecosystems and $CO_2$ Case Studies, Processes, and Perspectives, Ecological Studies*, vol. 187 (Springer, Berlin, 2006).
27. M. R. Shaw *et al.*, *Science* **298**, 1987 (2002).
28. E. A. Ainsworth, S. P. Long, *New Phytol.* **165**, 351 (2005).
29. A. D. B. Leakey, C. J. Bernacchi, F. G. Dohleman, D. R. Ort, S. P. Long, *Global Change Biol.* **10**, 951 (2004).
30. A. D. B. Leakey *et al.*, *Plant Physiol.* **140**, 779 (2006).
31. G. W. Wall *et al.*, *New Phytol.* **152**, 231 (2001).
32. M. J. Ottman *et al.*, *New Phytol.* **150**, 261 (2001).
33. M. R. Ashmore, in *Air Pollution and Plant Life*, J. N. B. Bell, M. Treshow, Eds. (Wiley, New York, 2002), pp. 89–118.
34. P. B. Morgan, T. A. Mies, G. A. Bollero, R. L. Nelson, S. P. Long, *New Phytol.* **170**, 333 (2006).
35. S. P. Long, E. A. Ainsworth, A. D. B. Leakey, P. B. Morgan, *Philos. Trans. R. Soc. London Ser. B* **360**, 2011 (2005).
36. J. S. Amthor, *Field Crops Res.* **73**, 1 (2001).
37. Materials and methods for meta-analyses are available as supporting material on *Science* Online. Full results from the meta-analyses summarized in Table 1 are presented in table S2 with references in appendix S1. $C_4$ crop yield responses to elevated [$CO_2$] are presented in table S3 with references in appendix S2.


38. This work was supported by the Illinois Council for Food and Agricultural Research, Archer Daniels Midland Company, U.S. Department of Agriculture, U.S. Department of Energy (grant DE-FG02-04ER63849), and Illinois Agricultural Experiment Station.


### Supporting Online Material

www.sciencemag.org/cgi/content/full/312/5782/1918/DC1
Materials and Methods
Tables S1 to S3
References
Appendices S1 and S2

1 March 2006; accepted 15 May 2006
10.1126/science.1114722

# Frictional Afterslip Following the 2005 Nias-Simeulue Earthquake, Sumatra

**Ya-Ju Hsu,[1]* Mark Simons,[1] Jean-Philippe Avouac,[1] John Galetzka,[1] Kerry Sieh,[1] Mohamed Chlieh,[1] Danny Natawidjaja,[2] Linette Prawirodirdjo,[3] Yehuda Bock[3]**

**Continuously recording Global Positioning System stations near the 28 March 2005 rupture of the Sunda megathrust [moment magnitude ($M_w$) 8.7] show that the earthquake triggered aseismic frictional afterslip on the subduction megathrust, with a major fraction of this slip in the up-dip direction from the main rupture. Eleven months after the main shock, afterslip continues at rates several times the average interseismic rate, resulting in deformation equivalent to at least a $M_w$ 8.2 earthquake. In general, along-strike variations in frictional behavior appear to persist over multiple earthquake cycles. Aftershocks cluster along the boundary between the region of coseismic slip and the up-dip creeping zone. We observe that the cumulative number of aftershocks increases linearly with postseismic displacements; this finding suggests that the temporal evolution of aftershocks is governed by afterslip.**

Slip on faults occurs as a combination of relatively continuous aseismic creep and transient slip events. These transient events occur as earthquakes radiating seismic waves, and also as aseismic events with characteristic time scales of days to years. A better understanding of the physical factors that control the relative amounts and location of seismic and aseismic slip is a key goal in the study of fault mechanics and in particular can affect assessments of regional seismic and tsunami hazards. After a large earthquake, postseismic deformation may result from earthquake-induced slip along the plate interface, commonly referred to as afterslip, and as viscoelastic relaxation in the volume surrounding the fault rupture (*1*–*3*). Thus, well-positioned postseismic observations can probe the mechanical properties of subduction megathrusts and the media that surround them.

Geodetic and seismological investigations suggest that typical subduction megathrust earthquakes involve fault rupture at depths between ~10 km and ~50 km, and that rupture all the way up to the trench is rare (*4*). However, evidence for slip on the shallowest portions of a megathrust has been notoriously difficult to evaluate. We commonly assume that seismic slip decreases in both up-dip and down-dip directions, presumably bounded by regions where frictional behavior of the fault does not support stick-slip (i.e., seismic) rupture (*5*).

[1]Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena, CA 91125, USA. [2]Research Center for Geotechnology, Indonesian Institute of Sciences, Bandung 40135, Indonesia. [3]Institute of Geophysics and Planetary Physics, University of California, San Diego, La Jolla, CA 92093, USA.

*To whom correspondence should be addressed. E-mail: yaru@gps.caltech.edu

These variations with depth are thought to result from the direct effect of temperature on the rheology of the plate interface or from indirect effects associated with metamorphism (*6*–*8*). In particular, the frictional behavior of the shallowest part of the megathrust is commonly assumed to be governed by clay minerals that promote rate-strengthening friction (*6*). Down-dip of the seismogenic zone, the megathrust presumably creeps continuously at approximately the plate convergence rate. In contrast to down-dip variations in seismogenic behavior, the underlying causes for along-strike variations remain enigmatic (*9*–*13*). Furthermore, what is actually happening along the shallowest portion of the megathrust is poorly known because of a lack of proximal geodetic observations in most areas. In contrast to most seismically active subduction zones, the presence of islands relatively close to the trench and above the seismogenic portions of the Sunda megathrust makes it possible to constrain coseismic, postseismic, and interseismic processes from geodetic monitoring and paleogeodetic studies (Fig. 1).

The 28 March 2005 $M_w$ 8.7 Nias-Simeulue earthquake resulted from rupture of the subduction megathrust off the shore of northwestern Sumatra, southeast of the 26 December 2004 $M_w$ 9.1 Aceh-Andaman rupture (*14*, *15*). In this region, the Indo-Australian plate converges obliquely toward the Sunda Block at a rate of ~57 mm/year (*16*) (Fig. 1, inset). This convergence is approximately partitioned into a 40 mm/year trench-normal component on the megathrust and a 25 mm/year component of dextral slip along the Sumatran Fault (*17*, *18*). The subduction megathrust off the shore of Sumatra has produced four earthquakes with magnitudes larger than 8 since 1797, including a $M$ ~8.5 event near Nias in 1861 (*19*, *20*) that occurred in the approximate region of the southern asperity of the 2005 event. Paleogeodetic and recent continuous Global Positioning System (cGPS) data from the Sumatran GPS Array (SuGAr) (*21*) as well as survey-mode GPS data suggest that the shallow portion of the megathrust up-dip of the Batu Islands (Fig. 1) is creeping during the interseismic period (*19*, *22*). The 2005 rupture occurred beneath the northern portion of SuGAr, permitting a record of both coseismic and postseismic deformation.

**The coseismic and postseismic slip model.** We use cGPS data spanning the first 11 months after the 2005 Nias-Simeulue earthquake, including nine SuGAr sites and the station SAMP, installed by the Indonesian National Surveying Agency (BAKOSURTANAL) (Fig. 1) (*15*, *23*). Three sites (LEWK, BSIM, and LHWA) were installed about a month before the Nias-Simeulue earthquake; three sites (PBLI, BITI, and BTHL) were installed 5 months after the earthquake; and three sites (PSMK, PTLO, and PBAI) just south of the Equator were installed in mid-2002, well before the Aceh-Andaman earthquake. In addition, 102 field measurements of vertical displacements from coral microatolls constrain the coseismic slip model. In the modeling of postseismic deformation, we use only the cGPS data.

The models considered here assume that all surface deformation is caused by slip on a dip-

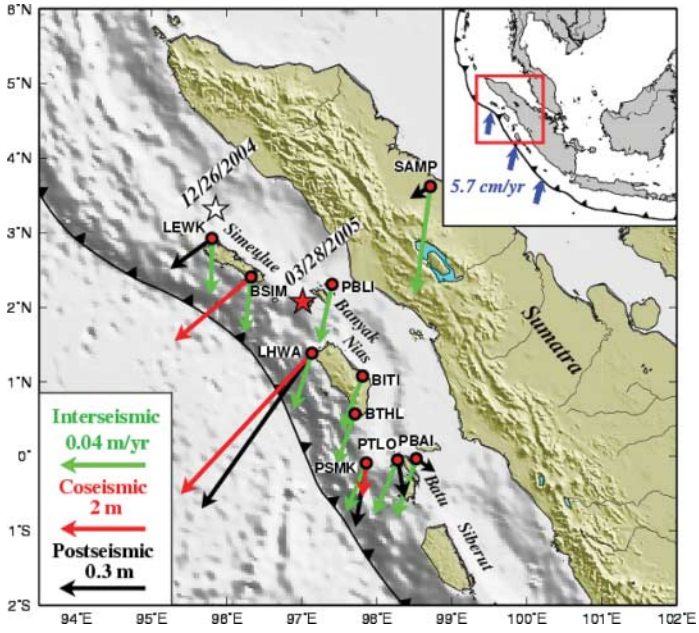



**Fig. 1.** Surface displacement estimated at the location of the cGPS stations in northern Sumatra for three time periods: The interseismic period before the 2004 Aceh-Andaman earthquake (green arrows), the 2005 Nias-Simeulue earthquake coseismic period (red arrows), and the 11-month period after the 2005 earthquake (black arrows). White and red stars indicate epicenters of 2004 Aceh-Andaman and 2005 Nias-Simeulue earthquakes, respectively. The black barbed line denotes the Sunda megathrust. Interseismic displacements for LEWK, BSIM, and LHWA are not derived from cGPS data, but rather from a model constrained by campaign GPS data (*46*). Estimates of coseismic displacement for PTLO and PBAI are too small to be visible at this scale. Three sites (PBLI, BITI, and BTHL) were installed after the earthquake and do not have estimates of coseismic displacements. (**Inset**) Regional geography with motion of the Australian plate relative to the Sunda Block indicated by blue arrows.

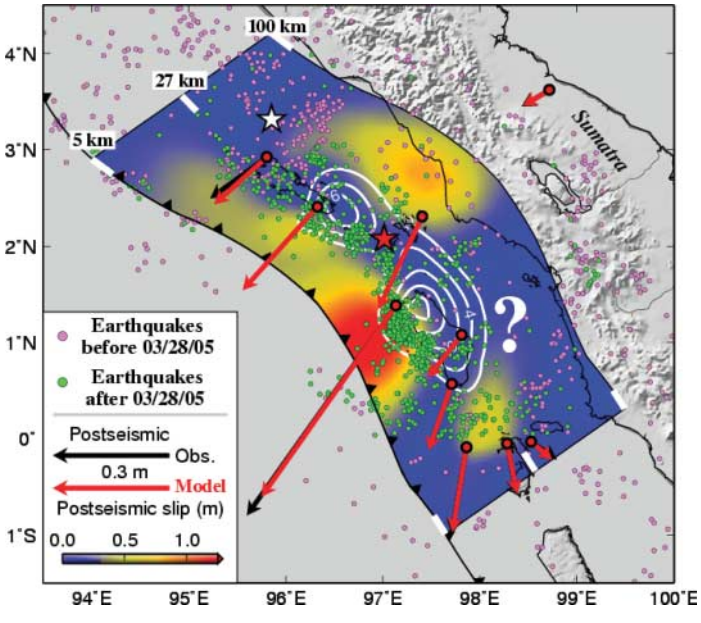



**Fig. 2.** Compilation of inferred coseismic and postseismic slip, illustrating extensive afterslip up-dip from the main shock and a lack of substantial overlap between seismogenic and aseismic regions. Distribution of coseismic slip is indicated by white contours at intervals of 2 m; color indicates cumulative postseismic slip during the 9 months after the main shock. Black and red vectors indicate observed and predicted GPS observations, respectively. Displacements measured at the stations deployed 5 months after the main shock are not shown. White and red stars are epicenters of 2004 Aceh-Andaman and 2005 Nias-Simeulue earthquakes, respectively. Pink and green dots denote earthquakes with $m_b$ > 4.5 before (*24*) and after (*43*) the 2005 event. The regions of high seismicity correspond to the transition between regions of coseismic and aseismic slip. The large question mark east of Nias indicates the region where afterslip may have occurred but is not detectable by the existing cGPS network. White tick marks on the northern and southern boundaries of the postseismic slip model indicate depths along the megathrust.

ping fault plane (i.e., the megathrust) embedded in a horizontally layered elastic space (*23*). The fault geometry differs slightly from that used previously (*24*); it is extended in length and width and has an additional segment at depth. The model fault approximately follows the curvature of the trench and changes dip from 10° to 30° at a depth of 27 km. The bottom of the model fault occurs at 100 km depth (Fig. 2) and follows the Wadati-Benioff zone, as defined by relocated seismicity spanning the period 1964 to 2005 (*25*). Constraints on the dip angle of the shallower portion come from the joint analysis of coseismic geodetic and seismic data (*26*), which yields a preferred dip angle for the seismogenic fault of between 8° and 12°. This range of dips can account simultaneously for the geodetic displacements, the seismic moment, and the amplitude of the spheroidal and toroidal normal modes (*26*).

We invert both coseismic and postseismic slip distributions with the use of the same fault geometry and a priori constraints on the rake (*23*). We have explored the impact of minimizing model roughness and model length (figs. S1 and S2). The time-dependent inversion of postseismic slip history adopts the extended network inversion filter (ENIF) approach (*27*, *28*). The GPS time series and model predictions, excluding estimates of benchmark wobble and reference-frame errors, are shown in fig. S3 (*29*).

The coseismic slip distribution, constrained by both the cGPS and the coral microatoll data (*24*), has two major loci of slip, separated by a region of negligible slip near the hypocenter (Fig. 2). Thus, the Nias-Simeulue earthquake resulted from two distinct ruptures: a $M_w$ 8.2 event to the north of the hypocenter and a $M_w$ 8.5 event to the south. Within the limits of resolution, the peak model slip values are 8 m (in the north) and 11 m (in the south). This coseismic model is similar to an earlier version that was determined using a slightly different megathrust geometry (*24*). Our present approach to damping is more rigorous and leads to a more spatially compact distribution of slip (*23*). The spatial coverage of data enables discrimination of the spatial extent of primary slip areas (fig. S1). The lack of up-dip coseismic slip is a robust feature of our model.

The postseismic slip distribution results from inversion of only the cGPS data. Even without microatoll data, resolution tests suggest that we can resolve up-dip afterslip well, whereas we can resolve down-dip afterslip only north of the hypocenter (fig. S2). The spatial distribution of postseismic slip remains basically stationary throughout the entire 11-month period (figs. S5 and S6). We note that the availability of data from three additional sites (PBLI, BITI, and BTHL) south of the rupture increases the model spatial resolution for later periods (fig. S6). Given that the pattern of slip is constant when we do not use these sites, we infer that the pattern we estimate including these sites is likely to be representative of the earlier period as well (fig. S6).

Afterslip occurs in three primary regions: up-dip, down-dip, and south of the coseismic rupture (Fig. 2). The most extensive resolvable region of afterslip lies up-dip of the coseismic rupture, where the maximum amount of afterslip over 11 months is ~1.4 m. Our confidence in the relative location of coseismic and postseismic slip is primarily driven by the observations of vertical displacements at two cGPS sites, LHWA and BSIM. These sites show coseismic uplift of 2.88 m and 1.60 m, respectively, but postseismic subsidence of 0.17 m and 0.05 m.

In the first 11 months after the earthquake, cumulative fault slip of ~0.5 m occurs both beneath the Batu Islands and down-dip of the coseismic rupture. The extent of down-dip slip is only loosely constrained because only one cGPS site, SAMP, provides data from mainland Sumatra. Indeed, there may have been considerable afterslip east of Nias, but we do not have sufficient data to resolve it. In many portions of the fault, the slip rate after 11 months is still about twice the long-term average plate convergence rate.

**Discussion.** The good fits to both vertical and horizontal components of coseismic and postseismic displacements suggest that our assumed fault geometry is adequate and that both coseismic and postseismic deformation resulted from slip on the megathrust. In the first 11 months of postseismic deformation, we see no evidence indicative of viscoelastic relaxation. This is probably due to the proximity of the geodetic stations to the rupture and to the short time span of the observations. Afterslip in the 11 months following the main shock has a geodetic moment of at least $2.5 \times 10^{21}$ N·m, equivalent to an $M_w$ 8.2 event or at least 25% of the coseismic moment. The estimate of postseismic geodetic moment is a minimum estimate, because there are likely to be regions of afterslip not resolved with the available data.

Most of the coseismic slip occurred within 150 km of the trench and within the locked fault zone, as inferred from modeling of paleogeodetic and geodetic data (*19*, *22*, *30*, *31*). To first order, and within the inherent limits of our model resolution, the region with afterslip surrounds the area that slipped during the earthquake. There appears to be little overlap between the coseismic and aseismic patches. Bearing in mind the limitations in spatial resolution, we find that 95% of the coseismic potency was released at depths between 13 and 48 km during the 2005 event, with a maximum at ~22 km (fig. S7).

Generally speaking, the inferred rakes of both coseismic and postseismic slip are parallel to each other, approximately perpendicular to the trench, and consistent with previous inferences of strain partitioning across the forearc (*32*, *33*). An exception to this behavior occurs

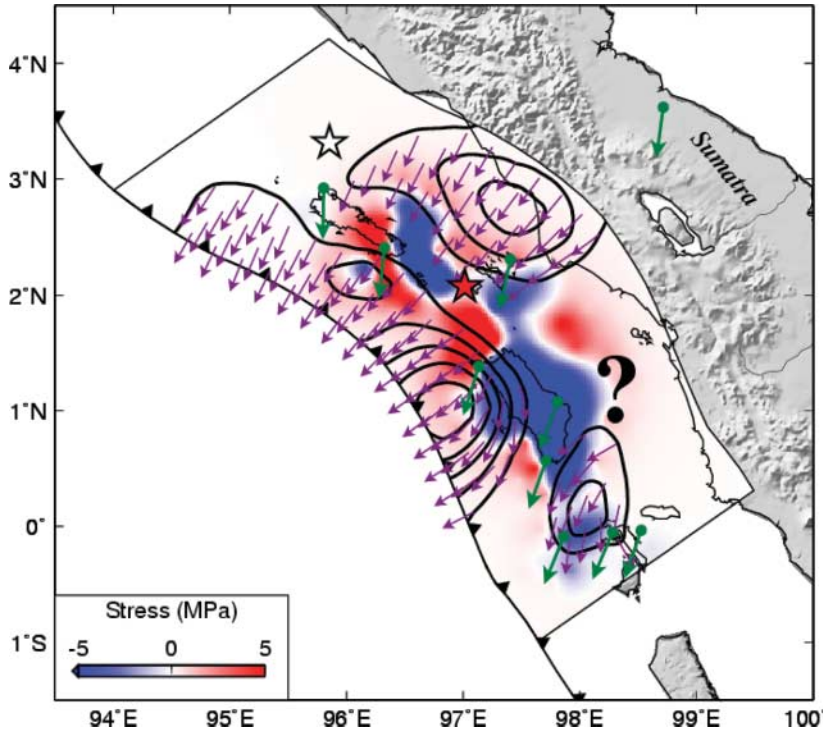



**Fig. 3.** Coulomb stress change Δ*CFS* (color) on the plate interface and amplitude of postseismic slip (black contours at 0.25-m intervals). For clarity, the model of coseismic slip (Fig. 2) is spatially smoothed before contouring. Purple arrows indicate the rake of the afterslip; green arrows indicate the rake of the interseismic velocity modeled/observed at the surface. White and red stars are epicenters of 2004 Aceh-Andaman and 2005 Nias-Simeulue earthquakes, respectively. The question mark east of Nias indicates a region where our afterslip model is not well constrained because of poor data coverage.

in the south, near the Batu Islands, where the coseismic and postseismic models predict slip rakes that are parallel to the general direction of convergence between the Australian plate and the Sunda Block (Fig. 3). Although poorly constrained, such behavior implies along-strike variations in strain partitioning.

The correlation between coseismic Coulomb stress change, $\Delta CFS$, on the plate interface and the distribution of afterslip (Fig. 3) (*23*) suggests that afterslip is a response of the megathrust to the sudden increase of stress due to the earthquake rupture. The patch with large afterslip up-dip of the coseismic rupture clearly coincides with a zone of increased $\Delta CFS$; this correlation also seems to hold for the down-dip afterslip zone, except east of Nias where the afterslip model is poorly resolved.

The coseismic slip distribution (Fig. 2) shows a distinct saddle between the Simeulue and Nias patches. This region of low slip corresponds to the location of the hypocenter and to the location of a distinct north-south bend in the band of aftershocks (Fig. 2). Previous studies have suggested that a structural tear occurs in this region, corresponding to the Batee Fault, and may be associated with more complex megathrust geometry than that used here (*17*, *34*).

To test the extent to which the spatial variations in slip are stationary in time, we consider historical and paleogeodetic records (*19*, *20*). Southeast of Nias, beneath the Batu Islands, the region of 2005 afterslip correlates with a prominent cluster of medium-sized earthquakes in the past century (*20*). However, coral microatolls show that, as in 2005, vertical deformation during the great historical earthquakes of 1797, 1833, and 1861 was small (*35*). Hence, we presume that megathrust slip in this region is primarily aseismic, with the medium-sized earthquakes representing small, spatially limited locked regions (*22*). We have inferred similarly low seismogenic coupling beneath the northern half of Simeulue in the region separating the 2004 and 2005 events (*24*). Seismicity preceding the 2005 earthquake (*25*) reveals distinct northeast-southwest–trending zones near northeastern Simeulue and near the Batu Islands (Fig. 2). We suggest that these zones that experience frequent small earthquakes during the interseismic period are likely to be regions dominated by aseismic slip. Because of the higher rates of seismicity in these regions, it may be more likely that a large megathrust event would nucleate there and eventually grow into areas that are more tightly coupled. Such behavior may be seen for the Nias-Simeulue earthquake as well as the 2004 Aceh-Andaman and 1995 Antofagasta, Chile, earthquakes (*12*, *15*).

The cGPS postseismic displacement histories are well fit if we assume that afterslip results from rate-strengthening frictional sliding of the plate interface in response to the coseismic stress change (Fig. 4, A to C) (*23*). The model is a system consisting of a spring and a slider with a single degree of freedom (*36*), where the slider obeys an experimental rate-strengthening friction law (*37*, *38*): $\tau_{ss} = \sigma_n\mu^* + A\sigma_n \ln(V/V^*)$, where $\tau_{ss}$ is the driving shear stress, $\sigma_n$ is the normal stress, $A$ is a positive rheological parameter, $V$ is the sliding velocity, and $\mu^*$ and $V^*$ are reference values. The postseismic displacement follows the predicted $\log[1 + (t/T_{GPS})]$ temporal evolution, where the characteristic time, $T_{GPS}$, is estimated to be ~3 days, and $d\tau_{ss}/d \ln V = A\sigma_n$ is on the order of 0.2 to 0.7 MPa. Alternatively, by analyzing the evolution of slip as a function of the evolving postseismic stresses according to our afterslip model (*39*), we find $A\sigma_n$ to be ~0.2 MPa both up-dip and down-dip of the coseismic rupture (fig. S8). If we assume that hydrostatic ambient pore pressure gives values of effective normal stress due to overburden of ~200 MPa and ~1000 MPa for the up-dip (~10 km) and down-dip regions (~60 km), respectively, these values result in estimates of $A$ of $\sim 5 \times 10^{-4}$, comparable to the value of $\sim 10 \times 10^{-4}$ at 35 km depth derived from afterslip following the 2003 $M_w$ 8.0 Tokachi-oki, Japan, earthquake (*39*), or the value of $3 \times 10^{-4}$ at 50 km depth for the 2001 $M_w$ 8.4 Arequipa, Peru, earthquake (*40*). These values are one to two orders of magnitude lower than estimated in laboratory studies (*41*, *42*). A value of $A$ at the lower end of experimental estimates ($\sim 50 \times 10^{-4}$) would imply a low effective normal stress of ~40 MPa. Any explanation for such a low value is conjectural; high pore pressure is one possibility.

About 2100 aftershocks with body wave magnitude $m_b > 3$ (*43*) occurred in the first year following the 2005 event. These aftershocks amount to only ~7% of the postseismic geodetic moment, indicating that afterslip was essentially aseismic. Most of these aftershocks form a distinct

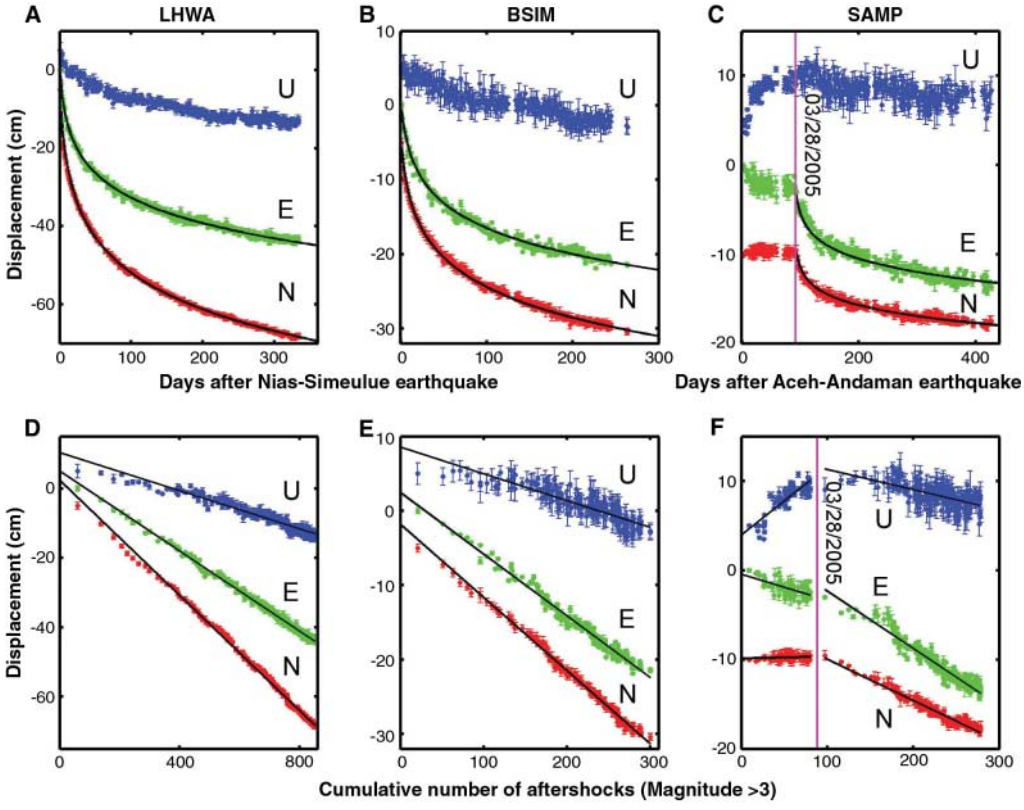



**Fig. 4.** Observed and modeled postseismic displacements (**A** to **C**) and the relationship of these displacements to the cumulative number of aftershocks near each of the stations (**D** to **F**). Black solid lines in (A) to (C) are estimated from a one-dimensional spring-slider model in which afterslip obeys a velocity-strengthening friction law [see (*23*) for analytical functions and model parameters] (*44*). Blue, green, and red refer to vertical (U), east (E), and north (N) displacements, respectively. Note that the scale differs between panels. Regions used to calculate cumulative seismicity are shown in fig. S10.

trench-parallel band between the coseismic slip patches and the up-dip zone of afterslip (Fig. 2). A less striking band of aftershocks northeast of Simeulue corresponds to the down-dip and southern edges of the Simeulue coseismic slip patch.

The temporal evolution of aftershocks may reflect a delayed response to the coseismic stress change of a population of small, volumetrically distributed, secondary faults (*42*). According to this model, the cumulative number of aftershocks should follow a $\log(1 + t/T_{as})$ evolution, mathematically equivalent to that of afterslip. However, we may expect different characteristic times, $T_{as}$ and $T_{GPS}$, because each process should reflect independent responses to the coseismic stress change. An alternative mechanism has seismicity controlled by the stressing rate associated with afterslip (*44*). In this case, both processes should have the same characteristic time, as was found for aftershocks of the 1999 Chi-Chi, Taiwan, and 2001 Arequipa, Peru, earthquakes (*40*, *44*). A third potential model has aftershocks driven by the coseismic stress change but also includes the spatial and temporal modulation of stress-associated afterslip (*42*).

To test these models, we consider two near-field cGPS sites (LHWA and BSIM), which we compare to the temporal evolution of nearby aftershocks, and one far-field cGPS site (SAMP), which we compare to the deep aftershocks. We find a clear linear relationship between cumulative displacement and cumulative number of aftershocks in regions near these cGPS sites (Fig. 4, D to F). For SAMP, this linear relationship exists both after the 2004 $M_w$ 9.1 event and after the 2005 event. The linear relationship between the cumulative number of aftershocks and the cumulative displacement at large times is not surprising (when presumably $t \gg T_{as}$ and $t \gg T_{GPS}$) because both processes should then vary linearly with the logarithm of time. The linear correlation in the early postseismic phase, when both aftershocks and afterslip depart from a linear dependence on the logarithm of time (Fig. 5 and fig. S9), is more meaningful. In the context of the functional form adopted here, the aftershocks and the afterslip have characteristic times that differ by no more than a factor of 2, although we cannot determine whether the times are exactly the same (Fig. 5). We note that a related observation for repeating earthquakes in the Loma Prieta aftershock zone suggests that the recurrence interval follows the characteristic inverse time decay, consistent with a model in which the repeating events are driven by creeping in the surrounding medium (*45*).

Our present models are limited by the lack of detailed structural information, in particular the role of splay and strike-slip faults and variations in elastic structure. Despite the limitations of the existing models, the 2005 Nias-Simeulue earthquake clearly illustrates the importance of aseismic slip on the shallowest portion of the megathrust. A wide variety of interseismic, coseismic, and postseismic data from the northern Sumatran forearc suggests that frictional behavior of the megathrust varies abruptly in space but not rapidly in time. Early near-field deformation following the 2005 Nias-Simeulue earthquake is dominated by afterslip on the megathrust, and both afterslip and aftershocks have the same functional dependence on time. Although temperature might be a key factor controlling regional upper and lower limits to seismogenic patches and down-dip variations of frictional properties, other factors must be called upon to explain the long-lived along-strike variations of the mode of slip seen here and in other subduction zones (*12*, *13*).

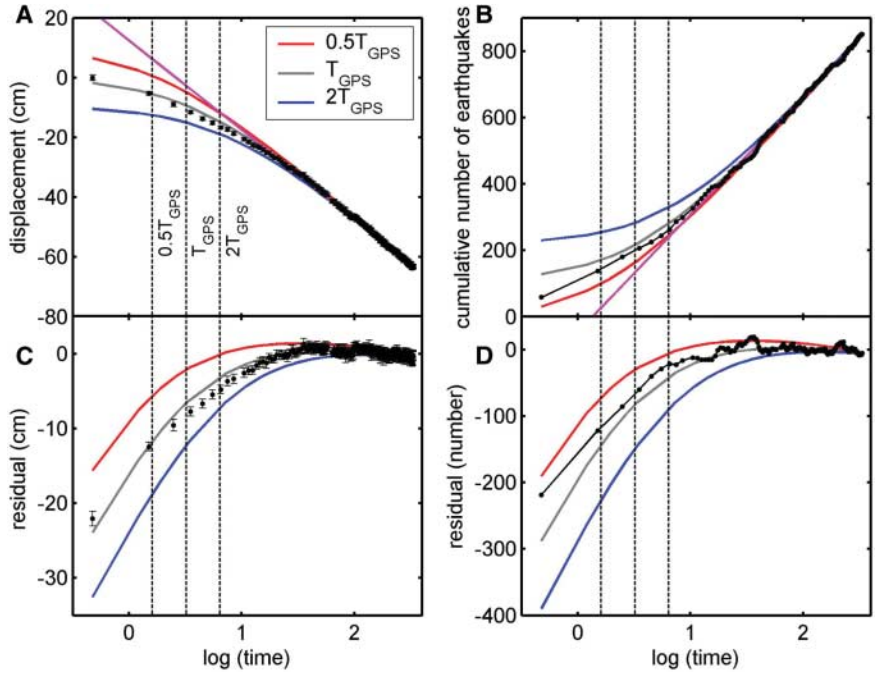



**Fig. 5.** (**A**) Comparison between the north component of postseismic displacement at LHWA (black dots, same as Fig. 4A) and modeled displacements [(*23*), eq. S1]. The gray line corresponds to the best-fit parameter values and is the same as in Fig. 4A, but in linear-log space. The blue and red lines correspond to fixing the estimates of $T_{GPS}$ to be a factor of 2 greater and less than the optimal values, respectively. Pink lines denote roughly linear variations of postseismic displacement with the logarithm of time when $t \gg T_{GPS}$ [(*23*), eq. S4]. (**B**) Same as (A) but for cumulative number of aftershocks. We fix $T_{as}$ to be the same as $T_{GPS}$ and only invert for the amplitude [(*23*), eq. S6]. (**C** and **D**) Residuals between model curves and observations with respect to the pink line. The temporal evolutions of afterslip and aftershock are bounded by model curves with characteristic time between $0.5T_{GPS}$ and $T_{GPS}$.

**References and Notes**


1. S. E. Barrientos, G. Plafker, E. Lorca, *Geophys. Res. Lett.* **19**, 701 (1992).
2. S. C. Cohen, *Adv. Geophys.* **41**, 134 (1999).
3. C. Zweck, J. T. Freymueller, S. C. Cohen, *J. Geophys. Res.* **107**, 10.1029/2001JB000409 (2002).
4. J. F. Pacheco, L. R. Sykes, C. H. Scholz, *J. Geophys. Res.* **98**, 14133 (1993).
5. C. H. Scholz, *Nature* **391**, 37 (1998).
6. R. D. Hyndman, M. Yamano, D. A. Oleskevich, *Island Arc* **6**, 244 (1997).
7. D. A. Oleskevich, R. D. Hyndman, K. Wang, *J. Geophys. Res.* **104**, 14965 (1999).
8. S. M. Peacock, R. D. Hyndman, *Geophys. Res. Lett.* **26**, 2517 (1999).
9. T.-R. A. Song, M. Simons, *Science* **301**, 630 (2003).
10. R. E. Wells, R. J. Blakely, Y. Sugiyama, D. W. Scholl, P. A. Dinterman, *J. Geophys. Res.* **108**, 10.1029/2002JB002072 (2003).
11. Y. Yamanaka, M. Kikuchi, *J. Geophys. Res.* **109**, 10.1029/2003JB002683 (2004).
12. M. E. Pritchard, M. Simons, *J. Geophys. Res.*, in press.
13. S. Miyazaki, P. Segall, J. Fukuda, T. Kato, *Geophys. Res. Lett.* **31**, 10.1029/2003GL019410 (2004).
14. T. Lay *et al.*, *Science* **308**, 1127 (2005).
15. C. Subarya *et al.*, *Nature* **440**, 46 (2006).
16. Y. Bock *et al.*, *J. Geophys. Res.* **108**, 10.1029/2001JB000324 (2003).
17. K. Sieh, D. Natawidjaja, *J. Geophys. Res.* **105**, 28295 (2000).
18. J. F. Genrich *et al.*, *J. Geophys. Res.* **105**, 28327 (2000).
19. D. H. Natawidjaja *et al.*, *J. Geophys. Res.* **109**, B04306 (2004).
20. K. R. Newcomb, W. R. McCann, *J. Geophys. Res.* **92**, 421 (1987).
21. Data are available at the Sumatran Plate Boundary Project site (www.tectonics.caltech.edu/sumatra).
22. K. Sieh, S. N. Ward, D. Natawidjaja, B. W. Suwargadi, *Geophys. Res. Lett.* **26**, 3141 (1999).
23. See supporting material on *Science* Online.
24. R. W. Briggs *et al.*, *Science* **311**, 1897 (2006).
25. E. R. Engdahl, A. Villasenor, H. R. DeShon, *Bull. Seismol. Soc. Am.*, in press.
26. O. Konca *et al.*, *Bull. Seismol. Soc. Am.*, in press.
27. J. J. McGuire, P. Segall, *Geophys. J. Int.* **155**, 778 (2003).
28. P. Segall, M. Matthews, *J. Geophys. Res.* **102**, 22391 (1997).
29. The fit of the model to the GPS time series is not ideal for a short period immediately after the main shock (fig. S4) because the ENIF underestimates the surface displacements.
30. L. Prawirodirdjo *et al.*, *Geophys. Res. Lett.* **24**, 2601 (1997).
31. M. Simoes, J. P. Avouac, R. Cattin, P. Henry, *J. Geophys. Res.* **109**, 10.1029/2003JB002958 (2004).
32. T. J. Fitch, *J. Geophys. Res.* **77**, 4432 (1972).
33. R. McCaffrey, *Geology* **19**, 881 (1991).

34. D. E. Karig, M. B. Lawrence, G. F. Moore, J. R. Curray, *J. Geol. Soc. (London)* **137**, 77 (1980).
35. D. Natawidjaja, thesis, California Institute of Technology (2003).
36. H. Perfettini, J. P. Avouac, *J. Geophys. Res.* **109**, 10.1029/2003JB002917 (2004).
37. J. H. Dieterich, *J. Geophys. Res.* **84**, 2161 (1979).
38. A. Ruina, *J. Geophys. Res.* **88**, 359 (1983).
39. S. Miyazaki *et al.*, *Geophys. Res. Lett.* **31**, 10.1029/2004GL021457 (2004).
40. H. Perfettini, J. P. Avouac, J. C. Ruegg, *J. Geophys. Res.* **110**, 10.1029/2004JB003522 (2005).
41. M. L. Blanpied, D. A. Lockner, J. D. Byerlee, *J. Geophys. Res.* **100**, 13045 (1995).
42. J. Dieterich, *J. Geophys. Res.* **99**, 2601 (1994).
43. Seismicity obtained from Incorporated Research Institutions for Seismology (IRIS) (www.iris.washington.edu/SeismiQuery/events.htm).
44. H. Perfettini, J. P. Avouac, *J. Geophys. Res.* **109**, 10.1029/2003JB002488 (2004).
45. D. P. Schaff, G. C. Beroza, B. E. Shaw, *Geophys. Res. Lett.* **25**, 4549 (1998).
46. M. Chlieh *et al.*, paper presented at the American Geophysical Union fall meeting, San Francisco, 5 to 9 December 2005, abstract A05j.
47. Supported by the Gordon and Betty Moore Foundation and the Indonesian International Joint Research Program (RUTI). We thank two anonymous reviewers for constructive comments, as well as R. W. Briggs, H. Perfettini, and A. J. Meltzner for valuable discussions. This is Caltech Tectonics Observatory contribution number 40 and Caltech Seismological Laboratory contribution number 9146.

**Supporting Online Material**

www.sciencemag.org/cgi/content/full/312/5782/1921/DC1
Materials and Methods
Figs. S1 to S10
References

2 March 2006; accepted 17 May 2006
10.1126/science.1126960

# REPORTS

# Spin-Wave Lifetimes Throughout the Brillouin Zone

**S. P. Bayrakci,[1]* T. Keller,[1,2] K. Habicht,[3] B. Keimer[1]**

We used a neutron spin-echo method with microelectron-volt resolution to determine the lifetimes of spin waves in the prototypical antiferromagnet $MnF_2$ over the entire Brillouin zone. A theory based on the interaction of spin waves (magnons) with longitudinal spin fluctuations provides an excellent, parameter-free description of the data, except at the lowest momenta and temperatures. This is surprising, given the prominence of alternative theories based on magnon-magnon interactions in the literature. The results and technique open up a new avenue for the investigation of fundamental concepts in magnetism. The technique also allows measurement of the lifetimes of other elementary excitations, such as lattice vibrations, throughout the Brillouin zone.

The concept of elementary excitations is one of the basic pillars of the theory of solids. In the low-temperature, long-wavelength limit, such excitations do not interact and have an infinite lifetime. For nonzero temperatures and momenta, the lifetimes of elementary excitations are generally limited by collisions with other excitations, with important consequences for the macroscopic properties of solids. For instance, the thermal expansion of solids can be understood as a consequence of collisions between lattice vibrations (phonons). Because of their comparatively simple Hamiltonians, magnetically ordered states are excellent testing grounds for theories of elementary excitations and their interactions. Nevertheless, the damping of spin waves in antiferromagnets has remained an open problem for four decades. Theoretical calculations of magnon lifetimes have been carried out since the 1960s, with intensive development occurring on several fronts in the early 1970s. However, these activities ground to a halt by the mid-1970s due to the lack of appropriate experimental data, namely, from momentum-resolved measurements with sufficient energy resolution. The only low-temperature data available were taken with momentum $q \cong 0$, in antiferromagnetic resonance (AFMR) and parallel pumping measurements (*1*, *2*). Because of the limited energy resolution, momentum-resolved data from neutron spectroscopy (*3*), by contrast, were confined to the critical regime extremely close to the Néel temperature ($T_N$), where most theories do not apply. Until recently, no other experimental techniques were available that permitted high-resolution measurements of excitation lifetimes at low temperatures over the whole Brillouin zone. We report on a neutron spectroscopy method with μeV resolution that is used to measure spin-wave (magnon) lifetimes in the prototypical antiferromagnet $MnF_2$ over the temperature range 0.04 to 0.6 $T_N$. The results subject long-standing theoretical predictions to a first experimental test and hold promise as a probe of elementary excitations in quantum

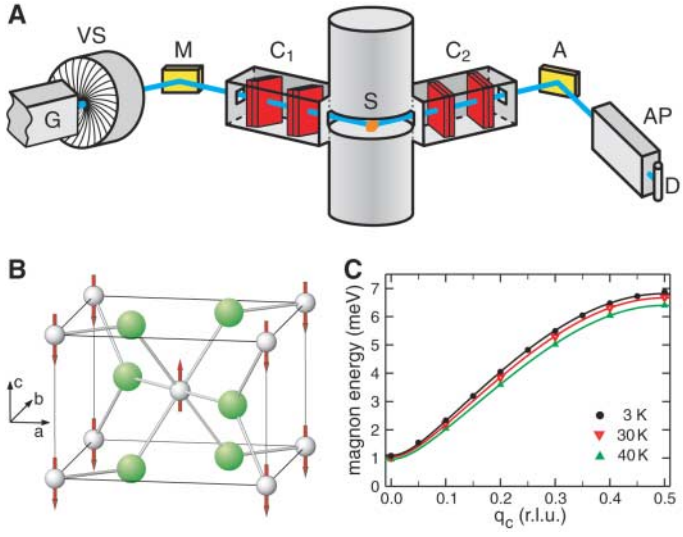



**Fig. 1.** (**A**) A diagram of the spectrometer TRISP at the FRM-II. G denotes the polarizing guide and AP the transmission polarizer; M and A are the monochromator and analyzer, as in TAS. S is the sample and D the detector; VS indicates the velocity selector. The resonance coil pairs ($C_1$ and $C_2$) are shown in red, and the mu-metal shielding boxes that enclose them in gray. The blue ray represents the path of the neutrons through the spectrometer, from left to right in the diagram. (**B**) The crystal and magnetic structure of $MnF_2$. The gray (smaller) spheres represent $Mn^{2+}$ ions and the green (larger) spheres the $F^-$ ions. The arrows indicate the relative directions of the $Mn^{2+}$ spins on the respective sublattices. (**C**) The magnon dispersion along the $q_c$ direction at three selected temperatures at and below 40 K. The data were taken on TRISP during the course of the linewidth measurements. The curves show the results of fits based on the same spin-wave result used by Okazaki *et al.* (*8*), in which the anisotropy is expressed by a single-ion form and the interactions of up to third-nearest neighbors are taken into account.

[1]Max-Planck-Institut für Festkörperforschung, Heisenbergstrasse 1, 70569 Stuttgart, Germany. [2]Forschungsreaktor München II, Zentrale Wissenschaftliche Einrichtung der Technischen Universität München (ZWE FRM II), Lichtenbergstrasse 1, 85748 Garching, Germany. [3]Hahn-Meitner-Institut, Glienickerstrasse 100, 14109 Berlin, Germany.

*To whom correspondence should be addressed. E-mail: bayrakci@fkf.mpg.de

magnets. The technique is also widely applicable to other elementary excitations, such as phonons and crystal-field excitations.

The determination of magnon lifetimes at low temperatures requires an energy resolution in the μeV range, about two orders of magnitude better than can be achieved by standard neutron triple-axis spectroscopy (TAS). We have obtained the requisite gain in resolution by manipulating the Larmor phase of the neutron spin with magnetic fields (*4*). The TRIple-axis resonance SPin echo spectrometer (TRISP) spectrometer (Fig. 1A) weds the capability of TAS of accessing collective excitations throughout the Brillouin zone to the extremely high energy resolution of neutron spin-echo spectroscopy (*5*). As in typical spin-polarized triple-axis spectrometry, the neutrons impinging upon the sample are polarized, and the polarization of neutrons scattered from the sample is measured. On TRISP, this is accomplished through the use of a polarizing neutron guide and a transmission polarizer, respectively. However, in analogy to neutron spin-echo spectrometry, the TRISP spectrometer also includes regions of effectively constant magnetic field that are produced by pairs of radio-frequency (RF) resonance coils inserted symmetrically (i) between the monochromator and sample and (ii) between the sample and the analyzer (*6*). The RF frequencies in the coils are tuned such that each detected neutron that creates an excitation lying on the magnon dispersion curve has the same net Larmor phase after traversing the two spin-echo arms, independent of small variations in the wave vector of the excitation. The neutron spin polarization determined at the detector is then a measure of the linewidth (inverse lifetime) of the magnon. In this way, the measured linewidth is decoupled (to first order) from the spread in energy of the neutrons incident on the sample, which contributes to the instrumental resolution in TAS. [For a detailed description of the technique, see the Materials and Methods (*7*).]

We chose the antiferromagnet $MnF_2$ for the experiment, because its magnetic ground state and excitations have been investigated extensively. The lattice structure and magnetic ordering of $MnF_2$ are shown in Fig. 1B. $MnF_2$ has the body-centered tetragonal structure, with $a = b = 4.87$ Å and $c = 3.30$ Å. The $Mn^{2+}$ ions have spin $S = 5/2$, and the spin in the center of the unit cell is oriented antiparallel to those at the corners. The strongest magnetic interaction is between second-nearest-neighbor $Mn^{2+}$ spins (corner and center spins) and is antiferromagnetic (*8*). A weaker, ferromagnetic interaction exists between nearest-neighbor spins (along the *c* axis). A relatively strong uniaxial anisotropy that is predominantly the result of dipole-dipole interactions (*9*, *10*) causes the spins to align along the *c* axis. $T_N$ is 67.6 K. The slope of the magnon dispersion is required to set the tilt angles of the RF coils and to determine the nonintrinsic contribution to the data (*7*). During the experiment, the spin-wave dispersion was therefore measured at each temperature at which linewidth data were taken; a partial data set is shown in Fig. 1C.

Figure 2 shows raw polarization data as a function of the spin-echo time τ. The spin-echo time is proportional to the frequency in the RF coils and the distance between the coils, and it also depends on the neutron wavelength. In a neutron spin-echo experiment, the dependence of the measured polarization on τ corresponds to the Fourier transform of the scattering function as a function of energy. The data in Fig. 2 are described well by an exponential decay, which indicates that the spectral function that characterizes the magnon linewidth is a Lorentzian in energy. The difference in linewidth (half-width at half maximum, or HWHM) between the upper two and lower two data sets is in each case only ~3 μeV, but it can be resolved clearly. The upper pair of data sets represents a difference in *q* of 0.05 r.l.u. at 15 K (*11*). For comparison, the HWHM of the corresponding TAS scans of the lower two magnons in Fig. 2, taken with fixed final neutron wave vector $k_f = 1.7$ Å$^{-1}$, is ~100 μeV.

The raw data were then corrected for instrumental and nonintrinsic effects (*7*). Figures 3 and 4 show the intrinsic magnon linewidth as a function of momentum *q* and temperature *T*, respectively. The linewidth generally increases with increasing *q* and *T*, due to the increasing likelihood of collisions with other excitations. However, Fig. 3 also shows that the linewidth deviates from this general trend and exhibits peaks as a function of *q* close to the center and the boundary of the antiferromagnetic Brillouin zone. The low-*q* peak is already present at 3 K, the lowest temperature covered by this experiment, and it evolves weakly with increasing temperature. This behavior is not described by the dominant magnon relaxation mechanisms for which quantitative predictions are available;

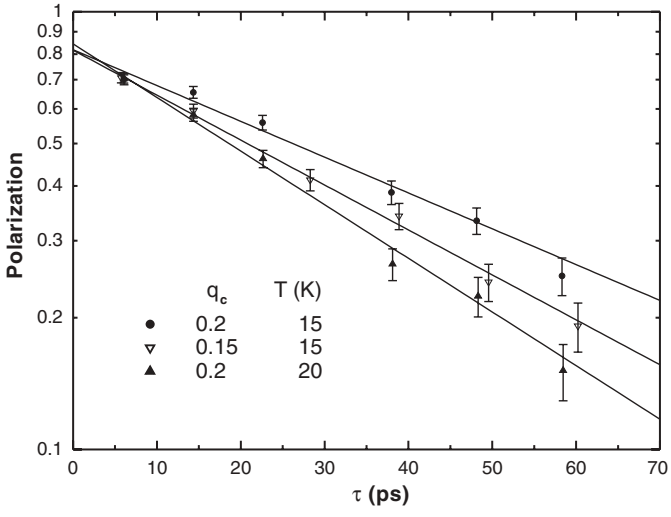



**Fig. 2.** Raw polarization data taken at (upper curve) $q = 0.2$ r.l.u. and $T = 15$ K, (middle curve) $q = 0.15$ r.l.u. and $T = 15$ K, and (lower curve) $q = 0.2$ r.l.u. and $T = 20$ K. The lines are exponential fits to the data. The corresponding Lorentzian magnon linewidths (HWHM) are 12.4 ± 0.8 μeV, 15.6 ± 0.9 μeV, and 18.6 ± 1.0 μeV, respectively.

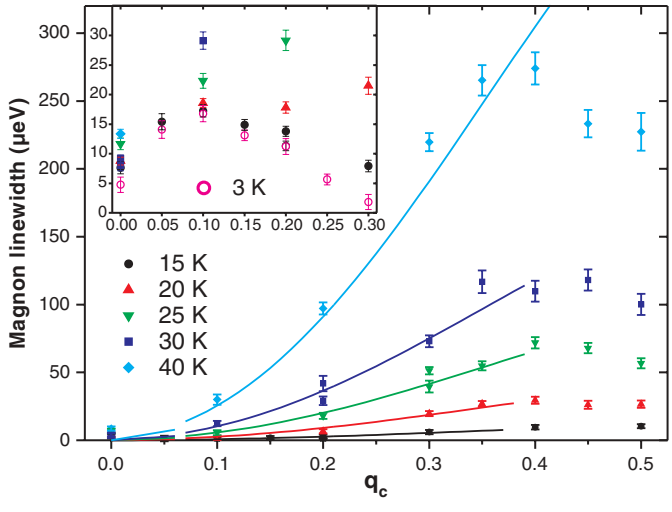



**Fig. 3.** Intrinsic magnon linewidth Γ at temperatures ranging from 15 to 40 K, as a function of *q*. We have plotted [Γ(*T*, *q*) − Γ(3 K, *q*)], where Γ(3 K, *q*) is given in the inset (see text). The curves show theoretical expressions from (*14*, *15*) (see text). Two different theoretical expressions are valid for the small-*q* case, depending on the magnitude of *q* relative to the anisotropy energy. Both expressions apply only to small *q*; Stinchcombe and Reinecke have applied one of them to data extending up to $q = 0.2$ r.l.u. for $MnF_2$ near $T_N$ (*15*). However, except at $q = 0$, the theory provides an excellent fit to the data as a function of *q* up to $q \sim 0.35$ r.l.u., both in the magnitude and in the *q* dependence.

possible origins will be discussed below. To facilitate comparison with these predictions, we have treated the 3-K linewidth data as a temperature-independent contribution and subtracted it from the higher-temperature data. The results are shown in Figs. 3 (main panel) and 4.

The intrinsic relaxation channel for magnons that has received by far the most attention in the literature is magnon-magnon scattering. In an "*n*-magnon" scattering event, a magnon (here, one excited by an incoming neutron) scatters off $(n/2 - 1)$ thermally excited magnons, producing $n/2$ scattered magnons that are in thermal equilibrium with the sample. In the absence of defects and external magnetic fields, the lowest-order interaction that limits the magnon lifetime is four-magnon scattering. Unfortunately, a comprehensive survey of the literature revealed very few theoretical predictions appropriate for comparison with our data, despite the existence of considerable work on four- and six-magnon interactions. This is either because the calculations used approximations valid only in high magnetic fields (for the purpose of comparison with AFMR data), or because strict inequalities that define the range of applicability of the theoretical results are extremely difficult to satisfy experimentally. An analytical expression was given by Harris *et al.* (*12*), who evaluated the contribution to the linewidth from four-magnon scattering processes analytically for the case of single-ion anisotropy with $q = 0$ (*13*). The corresponding result is shown in the bottom trace of Fig. 4. At low temperatures, the temperature dependence of the data is considerably weaker than that predicted by this theoretical result for $q = 0$. The best agreement of the magnitude occurs at 40 K, where the experimental result is ~30% larger than the theoretical one. Predictions for an anisotropy gap of dipolar origin, which would be more appropriate for $MnF_2$, are not available.

An additional relaxation channel, in which magnons are scattered by thermally excited longitudinal spin fluctuations, was considered by Stinchcombe and co-workers (*14*, *15*). The curves in Figs. 3 and 4 are based on this mechanism. (For $q = 0$, where the contribution of this relaxation is identically zero, we have shown the prediction of the four-magnon relaxation model, as discussed above.) For the larger-$q$ data, the linewidth far from $T_N$ is given approximately by

$$\Gamma_q(\mathrm{HWHM}) = \frac{\pi R_0' \rho^*}{4\mu^* R_0^2} q^* \varepsilon_q \times \frac{(1+\sigma)^2}{[1+\beta(1+\sigma)J(0)R_0']} \quad (1)$$

where $\varepsilon_q$ is the magnon energy, $q^* = 2\pi q/a$, $\mu^* = 2.969$ Å$^2$, $\rho^* = 5.864$ Å$^3$, and $\beta = 1/k_B T$, with $k_B$ the Boltzmann constant (*15*). The anisotropy parameter $\sigma$ is equal to 0.0184. The exchange parameter $J(0) = 6.02$ meV includes both first- and second-nearest-neighbor exchange interactions. The parameters $R_0$ and $R_0'$, which are both temperature dependent, can each be evaluated using either experimental data or results from mean-field theory (*15*), leading to considerable differences in the magnitude of the calculated linewidth and in its variation with temperature. In determining $R_0$, we used experimental data for the staggered magnetization (*16*). Calculation of $R_0$ from the Brillouin function produces linewidth values that agree at the lowest temperatures and begin to deviate with increasing temperature: At 40 K, the calculated linewidth is 11% smaller. For $R_0'$, we used the derivative of the Brillouin function. Calculation of $R_0'$ instead from experimental data for the parallel magnetic susceptibility (*17*) produces linewidth results that are 40% larger at 15 K and 30% smaller at 40 K.

Given the prominence of the magnon-magnon scattering channel in the literature, the excellent agreement between this model calculation and the experimental data is surprising. The dominance of the relaxation by longitudinal fluctuations is, however, consistent with arguments by Reinecke and Stinchcombe, who estimated that the contribution to the linewidth from four-magnon scattering is only $1/z$ of the magnitude of the above term (*18*, *19*). Here, $z$ is the number of neighbors that experience the strongest exchange interaction; $z = 8$ for $MnF_2$, for which case $z$ is the number of next-nearest neighbors. Because the analytical expression on which the curves in Figs. 3 and 4 are based is valid only at low $q$, deviation from the data at larger $q$ is not unexpected. The general expression for the linewidth resulting from scattering by longitudinal spin fluctuations (*14*, *15*) should be evaluated numerically at high $q$ to see if the peak as a function of $q$ can be reproduced (*20*, *21*).

Factors that might be expected to influence the linewidth at large $q$ include hybridization of the magnon mode with an optical phonon and the possible proximity of a two-particle continuum. The two lowest observed optical phonons along $(00l)$ lie at least 6 meV higher than the magnon mode over the entire Brillouin zone (*22*), so the former mechanism can be excluded. The data of Schweika *et al.* suggest that the single-magnon spectrum does not cross the two-particle continuum (*23*).

An explanation of the peak centered at $q \sim 0.1$ r.l.u. (Fig. 3, inset) requires a different mechanism (*24*). An additional potential source of linewidth is the hyperfine interaction, which gives rise to the scattering of electronic magnons from nuclear spin fluctuations (*25*). The contribution from the hyperfine interaction would only be weakly temperature dependent, because the nuclear spin system is already highly disordered thermally at 3 K. Four-magnon scattering terms in which one electronic and one nuclear magnon interact have indeed been shown to generate maxima in the linewidth at nonzero $q$, but estimates of the amplitude of this contribution are at least an order of magnitude smaller than the observed effect (*25*). The crossing of magnon and transverse acoustic phonon modes at $q \cong 0.04$ r.l.u. may also contribute to the peak (*22*, *26*, *27*). An additional relaxation mechanism that must be considered as a possible source of linewidth is that of magnon-phonon scattering. Experimental estimates of the linewidth due to magnon-phonon relaxation in $MnF_2$ in zero field range over three orders of magnitude, but again they appear too small to explain the observed peak (*28*–*30*). A theoretical estimate of the spin-lattice relaxation time (which should be of the same order of magnitude as the magnon-phonon relaxation times) corresponds to a linewidth of ~0.5 μeV at 25 K in $MnF_2$ (*31*, *32*). In this theory, the magnon-phonon interaction arises from the phonon modulation of the exchange interaction and is dominated by two-magnon–one-phonon processes. The result varies as $T^5$, which corresponds to a linewidth of 1 μeV at 30 K and 5 μeV at 40 K. The maximum potential contribution to our data would then be ~60% of the linewidth at $q = 0$ and 40 K.

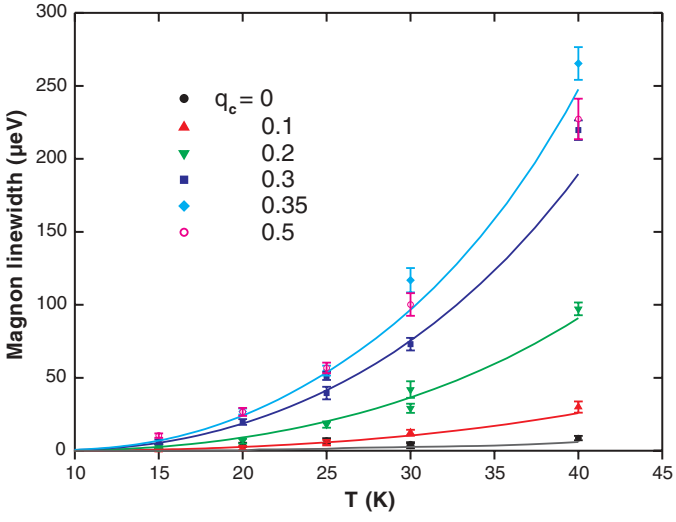



**Fig. 4.** Intrinsic magnon linewidth with $0 \leq q \leq 0.35$ r.l.u. and $q = 0.5$ r.l.u., shown as a function of temperature. As in Fig. 3, data taken at 3 K have been subtracted. For $0.1 \leq q \leq 0.35$ r.l.u., the curves show theoretical results from (*15*). For $q = 0$, results from Harris *et al.* are plotted (*12*, *13*).

Other mechanisms that may contribute to the presence of this peak include two- and three-magnon non–momentum-conserving processes that originate from scattering from defects (*25*, *33*, *34*). The linewidth originating from the latter process is peaked at intermediate $q$. Using parameters derived from comparison with data on $RbMnF_3$, one can estimate its contribution in $MnF_2$ to be two orders of magnitude smaller than the data (*34*).

The challenge to theory posed by the temperature- and momentum-dependent peaks in the magnon linewidth in $MnF_2$ should stimulate new activity in the field of spin-wave decay mechanisms. High-resolution lifetime measurements over the full Brillouin zone in a relatively simple antiferromagnet such as $MnF_2$ permit detailed evaluation of proposed processes, which should provide a basis for addressing such interactions in more complex magnetic systems.

### References and Notes


1. J. P. Kotthaus, V. Jaccarino, *Phys. Lett. A* **42**, 361 (1973).
2. J. Barak, S. M. Rezende, A. R. King, V. Jaccarino, *Phys. Rev. B* **21**, 3015 (1980).
3. M. P. Schulhof, R. Nathans, P. Heller, A. Linz, *Phys. Rev. B* **4**, 2254 (1971).
4. F. Mezei, *Z. Phys.* **255**, 146 (1972).
5. T. Keller *et al.*, *Appl. Phys. A* **74**, S332 (2002).
6. R. Golub, R. Gähler, *Phys. Lett. A* **123**, 43 (1987).
7. Materials and methods are available as supporting material on *Science* Online.
8. A. Okazaki, K. C. Turberfield, R. W. Stevenson, *Phys. Lett.* **8**, 9 (1964).
9. F. Keffer, *Phys. Rev.* **87**, 608 (1952).
10. J. Barak, V. Jaccarino, S. M. Rezende, *J. Magn. Magn. Mater.* **9**, 323 (1978).
11. We use a notation in which the momentum **Q** transferred by the neutron is $\mathbf{Q} = \mathbf{q} + \mathbf{G}$, where **q** is the momentum transfer within the Brillouin zone centered at the reciprocal lattice vector **G**. These quantities are expressed in reciprocal lattice units (r.l.u.). For instance, $\mathbf{Q} = (HKL)$, with $Q_a = H\,(2\pi/a)$, and $\mathbf{q} = (hkl)$, with $q_c = l(2\pi/c)$.
12. A. B. Harris, D. Kumar, B. I. Halperin, P. C. Hohenberg, *Phys. Rev. B* **3**, 961 (1971).
13. The conditions for application of the results of Harris *et al*. in various intervals of temperature and energy are quite restrictive due to the presence of strict inequalities, and the resulting analytical expressions for the linewidth are valid only in extremely limited regions, if at all. Here, we have treated these conditions as if they involved simple inequalities, and we have plotted the four solutions that apply to different temperature regions together.
14. M. G. Cottam, R. B. Stinchcombe, *J. Phys. C* **3**, 2305 (1971).
15. R. B. Stinchcombe, T. L. Reinecke, *Phys. Rev. B* **9**, 3786 (1974).
16. V. Jaccarino, in *Magnetism*, G. T. Rado, H. Suhl, Eds. (Academic Press, New York, 1965), vol. 2A, pp. 307–355.
17. S. Foner, in *Magnetism*, G. T. Rado, H. Suhl, Eds. (Academic Press, New York, 1963), vol. 1, pp. 383–447.
18. T. L. Reinecke, R. B. Stinchcombe, *Phys. Rev. B* **21**, 5198 (1980).
19. In contrast, in earlier work Cottam and Stinchcombe suggest that in the general case, the linewidth contribution from four-magnon scattering dominates at low temperatures (*14*).
20. A suggestion of nonmonotonic behavior in the linewidth as a function of $q$ is contained in (*14*). For the case in which the anisotropy takes the form of an anisotropy field $H_A$, the linewidth peaks at an intermediate value of $q$ and then declines as it approaches the Brillouin zone boundary at $q = 0.5$ r.l.u. However, the estimated value of $q$ at which the peak occurs would be only ~0.008 r.l.u. for $MnF_2$. The behavior of the linewidth at large $q$ for the case of anisotropic exchange was not evaluated.
21. Woolsey and White (*25*) predict a peak as a function of $q$ in $MnF_2$ for the linewidth due to a four-magnon scattering process in a relatively small magnetic field between 2 and 3 K. The peak becomes less pronounced and shifts to higher $q$ as the temperature increases to 3 K (from 0.30 r.l.u. at 2 K to 0.37 r.l.u. at 3 K). At a slightly higher field, only a simple maximum is present between 4 and 10 K for a four-magnon scattering process (*34*). The authors did not incorporate anisotropy in their calculations; later work by the same authors showed that the inclusion of anisotropy results in a decrease in the calculated linewidth of roughly an order of magnitude for $q = 0$ in a large magnetic field (*35*).
22. C. A. Rotter, cited by H. G. Smith, N. Wakabayashi, in *Dynamics of Solids and Liquids by Neutron Scattering*, S. W. Lovesey, Ed. (Springer-Verlag, Berlin, 1977), p. 93.
23. W. Schweika, S. V. Maleyev, T. Brückel, V. P. Plakhty, L.-P. Regnault, *Europhys. Lett.* **60**, 446 (2002).
24. The effect of the curvature of the magnon dispersion on the spin-echo resolution is largest at $q = 0$, and largest at low temperature, where the spin-wave stiffness is strongest. Thus, curvature not properly taken into account in the analytical correction calculation would produce an apparent linewidth. However, this spurious effect would be peaked at $q = 0$ and is therefore unlikely to be responsible for the observed peak.
25. R. B. Woolsey, R. M. White, *Int. J. Magn.* **2**, 51 (1972).
26. R. L. Melcher, *Phys. Rev. B* **2**, 733 (1970).
27. H. Montgomery, A. P. Cracknell, *J. Phys. C* **6**, 3156 (1973).
28. L. D. Rotter, W. D. Dennis, W. M. Yen, *Phys. Rev. B* **42**, 720 (1990).
29. G. J. Jongerden, A. F. M. Arts, J. I. Dijkhuis, H. W. de Wijn, *Phys. Rev. B* **40**, 9435 (1989).
30. L. D. Rotter, W. Grill, W. M. Dennis, *J. Lumin.* **45**, 130 (1990).
31. M. G. Cottam, *J. Phys. C* **7**, 2919 (1974).
32. A. Y. Wu, R. J. Sladek, *Phys. Rev. B* **25**, 5230 (1982).
33. R. Loudon, P. Pincus, *Phys. Rev.* **132**, 673 (1963).
34. R. B. Woolsey, thesis, Stanford University (1970).
35. R. M. White, R. Freedman, R. B. Woolsey, *Phys. Rev. B* **10**, 1039 (1974).


36. We thank G. Schmidt of the Crystal Growth Facility of the Cornell Center for Materials Research for the loan of a $MnF_2$ crystal of excellent quality, R. Henes and J. Major for the γ-ray diffractometry measurements, J. Peters for cryogenic assistance, G. Khaliullin and R. K. Kremer for illuminating discussions, P. Aynajian for participation in some of the calibration measurements, and R. Noack for technical assistance.


### Supporting Online Material

www.sciencemag.org/cgi/content/full/312/5782/1926/DC1
Materials and Methods
Fig. S1
References

22 March 2006; accepted 16 May 2006
10.1126/science.1127756

# Surface and Deep Ocean Interactions During the Cold Climate Event 8200 Years Ago

**Christopher R. W. Ellison,[1] Mark R. Chapman,[1]* Ian R. Hall[2]***

**Evidence from a North Atlantic deep-sea sediment core reveals that the largest climatic perturbation in our present interglacial, the 8200-year event, is marked by two distinct cooling events in the subpolar North Atlantic at 8490 and 8290 years ago. An associated reduction in deep flow speed provides evidence of a significant change to a major downwelling limb of the Atlantic meridional overturning circulation. The existence of a distinct surface freshening signal during these events strongly suggests that the sequenced surface and deep ocean changes were forced by pulsed meltwater outbursts from a multistep final drainage of the proglacial lakes associated with the decaying Laurentide Ice Sheet margin.**

Changes in the mode of operation of the Atlantic meridional overturning circulation (MOC) are thought to be an important driver of rapid, large-scale climate change (*1*, *2*) and are widely believed to be linked to changes in freshwater forcing (*3*). The climate of the present interglacial, which began about 10,000 years ago, is remarkably stable when compared to the last glacial period (*4*) with the exception of a single, large climatic excursion ~8200 years before the present (yr B.P.) (*5*), commonly referred to as the 8.2 ky event. It has been suggested that the short-lived cooling episode captured in the Greenland ice core records is one element of a broader climate signal that has an interhemispheric signature (*6*), but the magnitude and the sequence of changes surrounding this climate event are not always straightforward to decipher (*7*). The decay of the Laurentide Ice Sheet and the catastrophic final drainage of the large proglacial Lakes Agassiz and Ojibway at ~8470 yr B.P. (*8*) preceded the 8.2 ky event, so the precise relation between the lake outburst, the oceanic response, and the cooling observed in the wider circum-Atlantic region remains unresolved (*6*). However, the similarity in the timing of these two events, coupled with model-derived suggestions regarding the sen-

[1]School of Environmental Sciences, University of East Anglia (UEA), Norwich, NR4 7TJ, UK. [2]School of Earth, Ocean, and Planetary Sciences, Cardiff University, Main Building, Park Place, Cardiff, CF10 3YE, UK.

*To whom correspondence should be addressed. E-mail: Mark.Chapman@uea.ac.uk (M.R.C.); Hall@cardiff.ac.uk (I.R.H.)

sitivity of the MOC to freshwater inputs (*9*), have led to a widely held view that a meltwater-induced alteration to the ocean circulation may explain the trends observed in paleoclimate records. Despite recent efforts to detect MOC changes in paleoceanographic archives (*10*, *11*), clear and conclusive evidence has yet to be reported (*6*, *7*).

We analyzed deep-sea sediment core MD99-2251 recovered from the southern limb of the Gardar Drift in the subpolar North Atlantic (57°26.87′ N, 27°54.47′ W; 2620-m water depth), where the interaction of Iceland Scotland Overflow Water (ISOW) with the underlying sea-floor topography results in sediment focusing and exceptionally enhanced sediment accumulation rates. Calibrated accelerator mass spectrometry $^{14}$C dates (Fig. 1) (*12*) indicate a mean sediment accumulation rate of ~110 cm ky$^{-1}$ through the studied time interval of 9200 to 7200 years ago. Our sampling resolution over this interval is <20 years, with each sample representing an integrated signal of 6 to 10 years. Consequently, it has been possible to produce regional climate proxy records of unprecedented detail that reveal unambiguous information about the nature of subcentennial scale events and the relative phasing of surface and deep ocean changes. Because our climate proxy reconstructions are coregistered signals from within a single sediment core, the observed sequence of events and temporal offsets are robust features independent of correlation problems (*7*).

Variations in the relative abundance of the polar foraminifer *Neogloboquadrina pachyderma* sinistral (s.) coiling are routinely used to determine past positional changes of cold surface waters (*4*, *13*). The marked decline in *N. pachyderma* s. abundances from >60% to <2% at ~1600 cm in core MD99-2251 defines the northward retreat of polar waters at the onset of the Holocene (Fig. 1). Low abundances of the polar foraminifer (mean of 1.4%) persist throughout the Holocene with the exception of a marked excursion centered at ~1220 cm. This episode marks the cold climatic extreme in the Holocene, with faunal-derived transfer function estimates of summer sea surface temperature (SST) ~2°C colder than at present. The *N. pachyderma* s. percent abundance maximum is dated to 8290 yr B.P. in our record, in good agreement with the age of the pronounced cold excursion identified in the $\delta^{18}$O-derived atmospheric temperature record from the Greenland Ice Sheet Project 2 (GISP2) ice core (*5*, *14*). Although age estimates for the cold extreme differ by ~100 years in the oceanic and ice proxy records, this difference is not significant compared with the combined dating uncertainties (*12*, *15*), and we consider the surface ocean and atmospheric cooling events to be synchronous. This interpretation is supported by the structural similarity across the temperature excursion, i.e., the rates of change associated with the cooling into and warming after the event together with the near-identical duration for the event of ~70 years in both archives (Fig. 2).

Both of our surface ocean proxies, % *N. pachyderma* s. and the $\delta^{18}$O composition of the planktonic foraminifer *Globigerina bulloides*, reveal the existence of a separate, earlier climate event centered at 8490 yr B.P. (Fig. 2). This indicates that the prominent 8.2 ky cooling was not an isolated event but may be the culmination of a change initiated ~200 years earlier. These two marked cooling episodes occur within an interval of reduced SSTs that existed from ~8900 to 8000 yr B.P. (Fig. 3). This multicentennial cool phase, which is broadly consistent with the recent identification of a 400- to 600-year climate anomaly preceding the 8.2 ky event (*7*), signifies a longer-term change in the pattern of ocean heat transport and may reflect the cumulative impact of outflow from melting ice sheets (*16*). The dust supply to Greenland, inferred from variations in potassium content in the GISP2 ice core (*17*), suggests a similar pattern of polar atmospheric reorganization over this interval.

The abrupt initial cooling at 8490 yr B.P., which lasted ~80 years according to the % *N. pachyderma* s. data, was coincident with a ~0.6 per mil (‰) depletion in planktonic foraminiferal $\delta^{18}$O values (Fig. 2). The magnitude of this $\delta^{18}$O excursion increases further once temperature and ice volume effects are incorporated (Fig. 3). The resulting shift in $\delta^{18}O_{seawater}$ composition can only readily be explained by a substantial freshening of the surface ocean. This perturbation corresponds to a salinity reduction of 1 to 1.3, assuming a freshwater input with an end-member $\delta^{18}$O composition of between –35‰ and –20‰ (*18*). The changes in surface ocean proxy records 8490 years ago in the subpolar North Atlantic have no large-scale counterpart within the GISP2 temperature data but are hinted at in other paleoclimate proxy records (*17*, *19*), and, significantly, the timing of these coregistered signals is indistinguishable from the timing of the catastrophic meltwater release from Lakes Agassiz and Ojibway at ~8470 yr B.P. (*8*). This discharge event, whose combined volume has been estimated at 163,000 km$^3$ [or 5.2 × 10$^6$ m$^3$ s$^{-1}$ (Sv) if released in one year (*16*)], is the only identified potential freshwater source of sufficient magnitude to explain the presence of the pronounced $\delta^{18}O_{seawater}$ signal at the distal, open ocean location of core MD99-2251.

The more prominent cooling episode, associated with a further SST decrease of ~1.2°C centered at 8290 yr B.P. (Fig. 3), occurs ~200 years after the initial lake outburst and surface ocean cooling event. This event, correlative with the 8.2 ky event observed in Greenland ice, has a wider oceanic imprint and is registered elsewhere both in the North Atlantic (*4*, *20*) and Norwegian Sea (*13*, *21*). Notably, the initiation of this SST cooling phase is synchronous with a second and less pronounced freshening of the surface ocean. This signal may be related to a later, smaller freshwater outburst from the western Agassiz

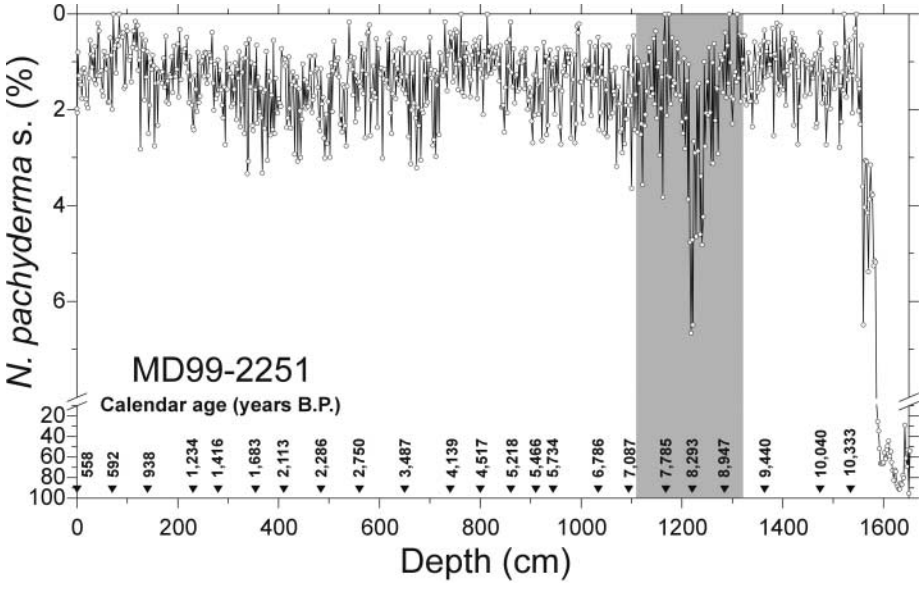



**Fig. 1.** The % abundance (>150 μm) of the polar foraminifer *N. pachyderma* s. coiling through the Holocene section of core MD99-2251 (note reversed axis). *N. pachyderma* s. % traces the rapid retreat of polar waters, marking the onset of the Holocene warming, at ~1600 cm and captures a significant, brief readvance of polar waters at ~1220 cm. The core chronology has its basis in 23 accelerator mass spectrometry (AMS) $^{14}$C dates that have been converted to calendrical ages by using the CALIB program (*31*). Position and calibrated ages of AMS $^{14}$C dates are indicated along depth axis (black triangles) (*12*). The shaded area highlights the study interval shown in Figs. 2 and 3 (i.e., 9200 to 7200 years ago). Mean sediment accumulation rate through this interval is ~110 cm ky$^{-1}$.

region that postdates the main Lake Agassiz and Ojibway discharge (*16*). The possibility of complex variations in freshwater forcing through changes in drainage pathways, discharge rate, reservoir refilling, and multipulse flooding has been highlighted previously (*22*). It is interesting to note that the two-step drawdown model (*16*), whereby the Lakes Agassiz and Ojibway meltwaters discharge in separate 113,100 km$^3$ (3.6 Sv) and 49,100 km$^3$ (1.6 Sv) pulses (for a period of 1 year), appears to be consistent with the multievent sequence found in our paleoceanographic records.

The impacts of these changes to the near-surface environment on the flow regime of deep ocean currents are determined by analyzing the grain size distributions of the terrigenous fine fraction (*23*). Variations in the mean size of the sortable silt ($\overline{SS}$) indicate that the near-bottom flow speed of ISOW, an important component of the Atlantic MOC, changed significantly during the 8.2 ky event (Fig. 3). The cooling and freshening of the surface ocean at 8490 yr B.P. coincides with the first signs of a longer-term reduction in the flow speed of ISOW. We document the most pronounced reduction in ISOW current vigor at the conclusion of the second surface ocean perturbation around 8290 yr B.P., with a rapid drawdown in ISOW flow speed taking place over a few decades. The minimum flow speeds attained at 8260 yr B.P. then remain subdued for a further 100 to 200 years. The stepped warming trends after the 8.2 ky event evident in the % *N. pachyderma* s. and GISP2 $\delta^{18}O$ records appear to be in phase and lead the recovery in the deep flow speeds by ~200 years, demonstrating a nonlinearity in the response of the MOC to such perturbations. The longer-term response of the deep ocean circulation to the 8.2 ky event indicated by the $\overline{SS}$ data is of similar duration to that estimated in modeling experiments (*24*). Both SST and salinity remained at intermediate values during the $\overline{SS}$ minimum, most likely sustained by a combination of the enhanced baseline meltwater discharge that persisted for centuries after the initial outburst (*16*) and the MOC adjustment time required to account for the altered surface ocean buoyancy forcing. SST and paleocurrent speed proxies are in phase by the time of the later climate transition at ~7900 yr B.P., with elevated flow speeds suggesting a somewhat enhanced MOC activity that may be analogous to the "climate overshoot" oceanic response associated with abrupt interstadial warming during the last glacial period (*18*).

We can interpret the decline in $\overline{SS}$ values across the 8.2 ky event as resulting from reduced deep water convection in the Nordic and Arctic seas in response to freshwater perturbations and the subsequent propagation of this signal to the deep ocean through ISOW flow rates. However, the vigor of ISOW recorded in MD99-2251 also may be modified by the degree of entrainment of surface and intermediate water masses in the Iceland Basin during its descent to depth after its overflow of the Iceland-Scotland Ridge. Such changes are controlled by the density contrast between the overflowing waters and the upper water column (*25*, *26*). Hence, the ISOW flow signal we observed may reflect a "shoaling" of intermediate waters. If the boundary between intermediate and deep waters rose in a similar fashion to the situation at the last glacial maximum (*27*), then the ISOW flow strength recorded at the MD99-2251 core site could be reduced. The Holocene perturbation in deep flow speed is ~50% of the relative magnitude of the circulation decrease that marked the end of the last interglacial period (~118 kyr B.P.) on the Gardar Drift (*28*), although such comparisons are only

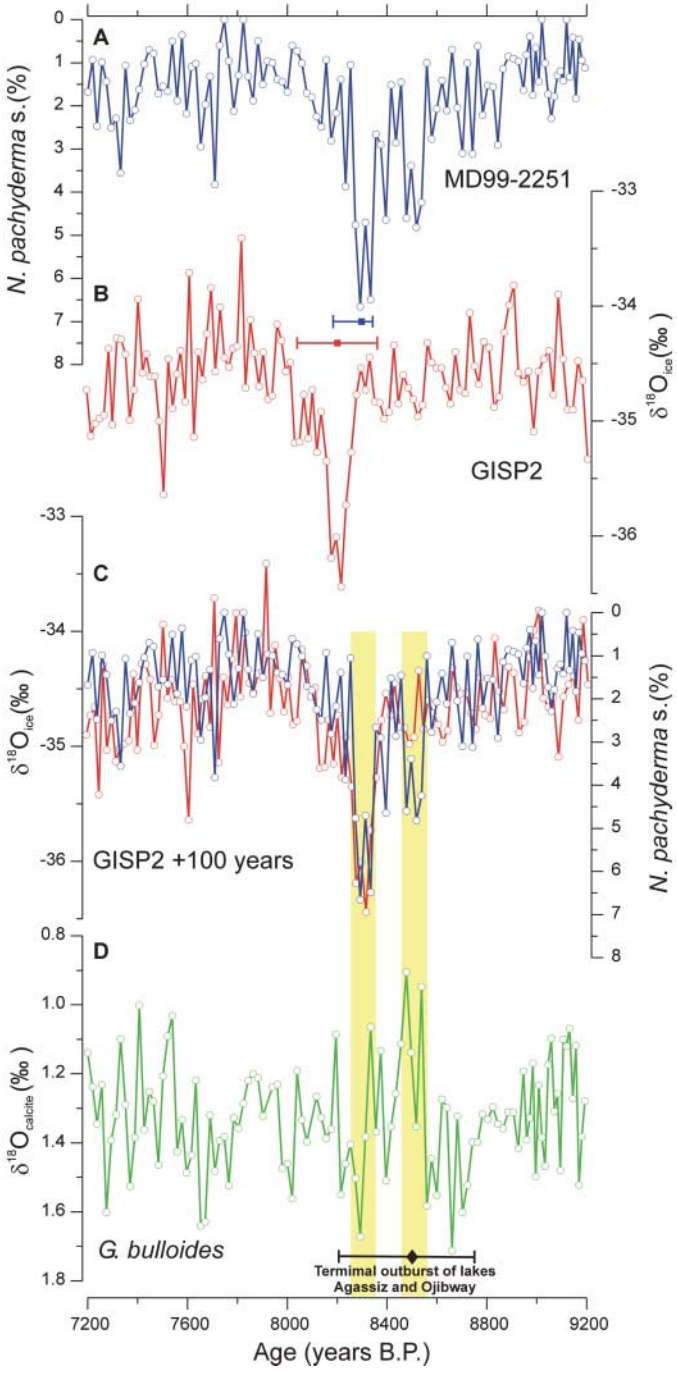



**Fig. 2.** Surface ocean climate proxy records in core MD99-2251 compared to the GISP2 ice core data over the interval 9200 to 7200 years ago. (**A**) *N. pachyderma* s. % abundance data and (**B**) the GISP2 temperature ($\delta^{18}O_{ice}$) record (*14*). Dating errors around the 8.2 ky event are shown for each time series. The lower and upper 2$\sigma$ limits on the calibrated AMS $^{14}C$ date for the cooling event in core MD99-2251 (8290 yr B.P. at 1220-cm depth) are 8180 and 8340 yr B.P., respectively. The GISP2 time scale has an estimated error of up to 2% on the basis of annual layer counting (*15*); this equates to a maximum dating uncertainty of ±164 years for 8200 years ago. (**C**) GISP2 $\delta^{18}O_{ice}$ record (red) overlain on the *N. pachyderma* s. % abundance data (blue) in MD99-2251, with the age scale of the GISP2 $\delta^{18}O_{ice}$ offset by 100 years (i.e., showing the interval 9100 to 7100 ice core years). The application of a constant age offset of +100 years, which is well within the dating uncertainties of either one of the two records, yields a good agreement between the cold extremes represented in each proxy record. (**D**) Planktonic foraminiferal stable isotope ($\delta^{18}O_{calcite}$) data measured for the species *G. bulloides*. Diamond marker and error bar indicates the timing of Lakes Agassiz and Ojibway outburst at ~8470 yr B.P. (*8*) and is associated with the first major shift in MD99-2251 proxy data (indicated by older shaded bar). Younger shaded bar highlights the largest cooling in the MD99-2251 proxy data, correlated with the 8.2 ky event.

meaningful if the prevailing MOC configurations were similar for both events. Whatever the origin of the ISOW flow signal, it is clear that we observe a significant change in the initial stages of the MOC. These findings lend support to model experiments that predict a reduction in MOC intensity during the 8.2-ky event (*6*, *9*, *24*).

Although two minima in solar output inferred from cosmogenic $^{14}C$ production records (*29*) may correlate with the key meltwater events and SST minima that we identify at 8490 and 8290 yr B.P., there is no clear relation between oceanic proxy records and solar variability through the remainder of the study interval. Hence, any causal linkage would appear to be dependent on other factors, such as the preconditioning of the North Atlantic by enhanced meltwater input. Rather, our coregistered paleoclimate records most likely demonstrate a cause and effect relationship: The 8.2 ky event was forced by the freshwater anomaly associated with the pulsed drainage of Lakes Agassiz and Ojibway, which caused extensive cooling and freshening in the North Atlantic Ocean, and this in turn can be directly linked to alterations in the MOC. Although the duration of the meltwater release is likely to be on the order of years (*16*) and the surface ocean perturbations seen in the subpolar North Atlantic are on the order of decades, the associated change in the deep ocean flow lasted ~400 years. Such information on the sensitivity and response of the climate system to a disturbance of the MOC during interglacial times is crucial and relevant to our possible future, because many global climate models suggest that an interruption of the MOC is a likely outcome of future climate change (*30*).

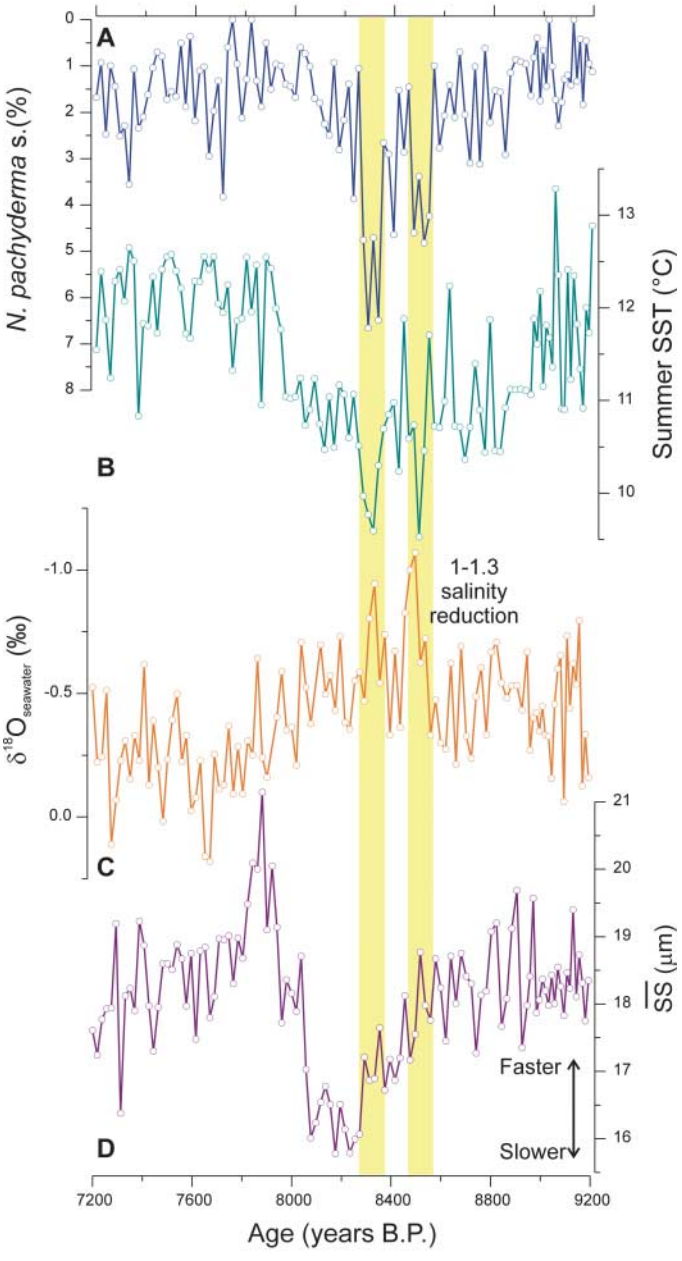



**Fig. 3.** Surface and deep ocean linkages over the interval from 9200 to 7200 years ago. (**A**) *N. pachyderma* s. % abundance data, (**B**) summer SST, (**C**) surface ocean stable isotope ($\delta^{18}O_{seawater}$) composition, and (**D**) $\overline{SS}$ mean grain size. SST values (error ± 1°C) were estimated by using the modern analog technique and are derived from full planktonic foraminiferal assemblage counts (~400 individuals) using the >150-μm size fraction (*12*). The depletion in $\delta^{18}O_{seawater}$ values at 8490 and 8330 yr B.P. correspond to reductions in surface ocean salinity of 1 to 1.3 and 0.5 to 0.7, respectively, assuming a freshwater end-member $\delta^{18}O$ composition of between –35‰ and –20‰ (*18*). The $\overline{SS}$ is a paleocurrent flow speed proxy (double arrow), where a higher mean reflects stronger flow of the depositing current and vice versa (*23*). Shaded bars are as in Fig. 2 and indicate the two pulses of cooling and freshening of surface ocean conditions that appear to correlate with a larger (3.6 Sv) meltwater outburst at ~7700 $^{14}C$ yr B.P. and a second, smaller meltwater discharge (1.6 Sv) dated at ~7600 $^{14}C$ yr B.P. (*16*).


**References and Notes**

1. J. F. McManus, R. Francois, J.-M. Gherardi, L. Keigwin, S. Brown-Leger, *Nature* **428**, 834 (2004).
2. S. Rahmstorf, *Nature* **419**, 207 (2002).
3. P. U. Clark, N. G. Pisias, T. F. Stocker, A. J. Weaver, *Nature* **415**, 863 (2002).
4. G. Bond *et al.*, *Science* **278**, 1257 (1997).
5. R. B. Alley *et al.*, *Geology* **25**, 483 (1997).
6. R. B. Alley, A. M. Agustsdottir, *Quat. Sci. Rev.* **24**, 1123 (2005).
7. E. J. Rohling, H. Pälike, *Nature* **434**, 975 (2005).
8. D. C. Barber *et al.*, *Nature* **400**, 344 (1999).
9. E. Bauer, A. Ganopolski, M. Montoya, *Paleoceanography* **19**, 10.1029/2004PA001030 (2004).
10. D. W. Oppo, J. F. McManus, J. L. Cullen, *Nature* **422**, 277 (2003).
11. I. R. Hall, G. G. Bianchi, J. R. Evans, *Quat. Sci. Rev.* **23**, 1529 (2004).
12. Materials and methods are available as supporting material on *Science* Online.
13. D. Klitgaard-Kristensen, H. P. Sejrup, H. Haflidason, S. Johnsen, M. Spurk, *J. Quat. Sci.* **13**, 165 (1998).
14. P. M. Grootes, M. Stuiver, J. W. C. White, S. Johnsen, J. Jouzel, *Nature* **366**, 552 (1993).
15. D. A. Meese *et al.*, *Science* **266**, 1680 (1994).
16. J. T. Teller, D. W. Leverington, J. D. Mann, *Quat. Sci. Rev.* **21**, 879 (2002).
17. S. R. O'Brien *et al.*, *Science* **270**, 1962 (1995).
18. M. A. Maslin, N. J. Shackleton, U. Pflaumann, *Paleoceanography* **10**, 527 (1995).
19. U. von Grafenstein, H. Erlenkeuser, A. Brauer, J. Jouzel, S. Johnsen, *Science* **284**, 1654 (1999).
20. M. Moros *et al.*, *Quat. Sci. Rev.* **23**, 2113 (2004).
21. B. Risebrobakken, E. Jansen, C. Andersson, E. Mjelde, K. Hevroy, *Paleoceanography* **18**, 10.1029/2002PA000764 (2003).
22. P. U. Clark *et al.*, *Science* **293**, 283 (2001).
23. I. N. McCave, B. Manighetti, S. G. Robinson, *Paleoceanography* **10**, 593 (1995).
24. H. Renssen, H. Goosse, T. Fichefet, *Paleoceanography* **17**, 1020 (2002).
25. G. G. Bianchi, I. N. McCave, *Nature* **397**, 515 (1999).
26. J. F. Price, M. O. Baringer, *Prog. Oceanogr.* **33**, 161 (1994).
27. J. C. Duplessy *et al.*, *Paleoceanography* **3**, 343 (1988).
28. I. R. Hall, I. N. McCave, M. R. Chapman, N. J. Shackleton, *Earth Planet. Sci. Lett.* **164**, 15 (1998).
29. R. Muscheler, J. Beer, M. Vonmoos, *Quat. Sci. Rev.* **23**, 2101 (2004).
30. R. A. Wood, M. Vellinga, R. Thorpe, *Philos. Trans. R. Soc. London Ser. A* **361**, 1961 (2003).
31. M. Stuiver *et al.*, *Radiocarbon* **40**, 1041 (1998).


32. We thank G. Bianchi, Cardiff, for running the mass spectrometer and H. Medley, Cardiff, and S. Bennett, UEA, for providing invaluable laboratory assistance. Financial support was provided by the Natural Environment Research Council (NERC) and NERC Radiocarbon Laboratory. Data are available from World Data Center-A for Paleoclimatology (www.ncdc.noaa.gov/paleo/data.html).


**Supporting Online Material**

www.sciencemag.org/cgi/content/full/312/5782/1929/DC1
Materials and Methods
Table S1
References


9 March 2006; accepted 16 May 2006
10.1126/science.1127213

# Phase Modifiers Promote Efficient Production of Hydroxymethylfurfural from Fructose

Yuriy Román-Leshkov, Juben N. Chheda, James A. Dumesic*

Furan derivatives obtained from renewable biomass resources have the potential to serve as substitutes for the petroleum-based building blocks that are currently used in the production of plastics and fine chemicals. We developed a process for the selective dehydration of fructose to 5-hydroxymethylfurfural (HMF) that operates at high fructose concentrations (10 to 50 weight %), achieves high yields (80% HMF selectivity at 90% fructose conversion), and delivers HMF in a separation-friendly solvent. In a two-phase reactor system, fructose is dehydrated in the aqueous phase with the use of an acid catalyst (hydrochloric acid or an acidic ion-exchange resin) with dimethylsulfoxide and/or poly(1-vinyl-2-pyrrolidinone) added to suppress undesired side reactions. The HMF product is continuously extracted into an organic phase (methylisobutylketone) modified with 2-butanol to enhance partitioning from the reactive aqueous solution.

We are entering an era of diminishing availability of petrochemical resources used to produce the energy and chemical materials needed by society. Abundant biomass resources are a promising alternative for the sustainable supply of valuable intermediates (such as alcohols, aldehydes, ketones, and carboxylic acids) to the chemical industry for production of drugs and polymeric materials. In this context, the high content of oxygenated functional groups in carbohydrates—the dominant compounds in biomass—is an advantage, in contrast to the drawbacks of such functionality for the conversion of carbohydrates to fuels. However, efficient processes must be developed for the selective removal of excess functional groups and the modification of others to create the desired products.

Here, we present a strategy for the selective dehydration of fructose to produce HMF, thereby providing a cost-effective route for the synthesis of this valuable chemical intermediate. Indeed, HMF and its ensuing 2,5-disubstituted furan derivatives can replace key petroleum-based building blocks (*1*). For example, HMF can be converted to 2,5-furandicarboxylic acid (FDCA) by selective oxidation, and Werpy and Petersen (*2*) and Pentz (*3*) have suggested that FDCA can be used as a replacement for terephthalic acid in the production of polyesters such as polyethyleneterephthalate (*2*) and polybutyleneterephthalate. They have also suggested that the reduction of HMF can lead to products such as 2,5-dihydroxymethylfuran and 2,5-bis(hydroxymethyl)tetrahydrofuran, which can serve as alcohol components in the production of polyesters, thereby leading to completely biomass-derived polymers when combined with FDCA. In addition, HMF can serve as a precursor in the synthesis of liquid alkanes to be used, for example, in diesel fuel (*4*).

Unfortunately, as noted by various authors (*5*–*8*), the industrial use of HMF as a chemical

Department of Chemical and Biological Engineering, University of Wisconsin, Madison, WI 53706, USA.

*To whom correspondence should be addressed. E-mail: dumesic@engr.wisc.edu

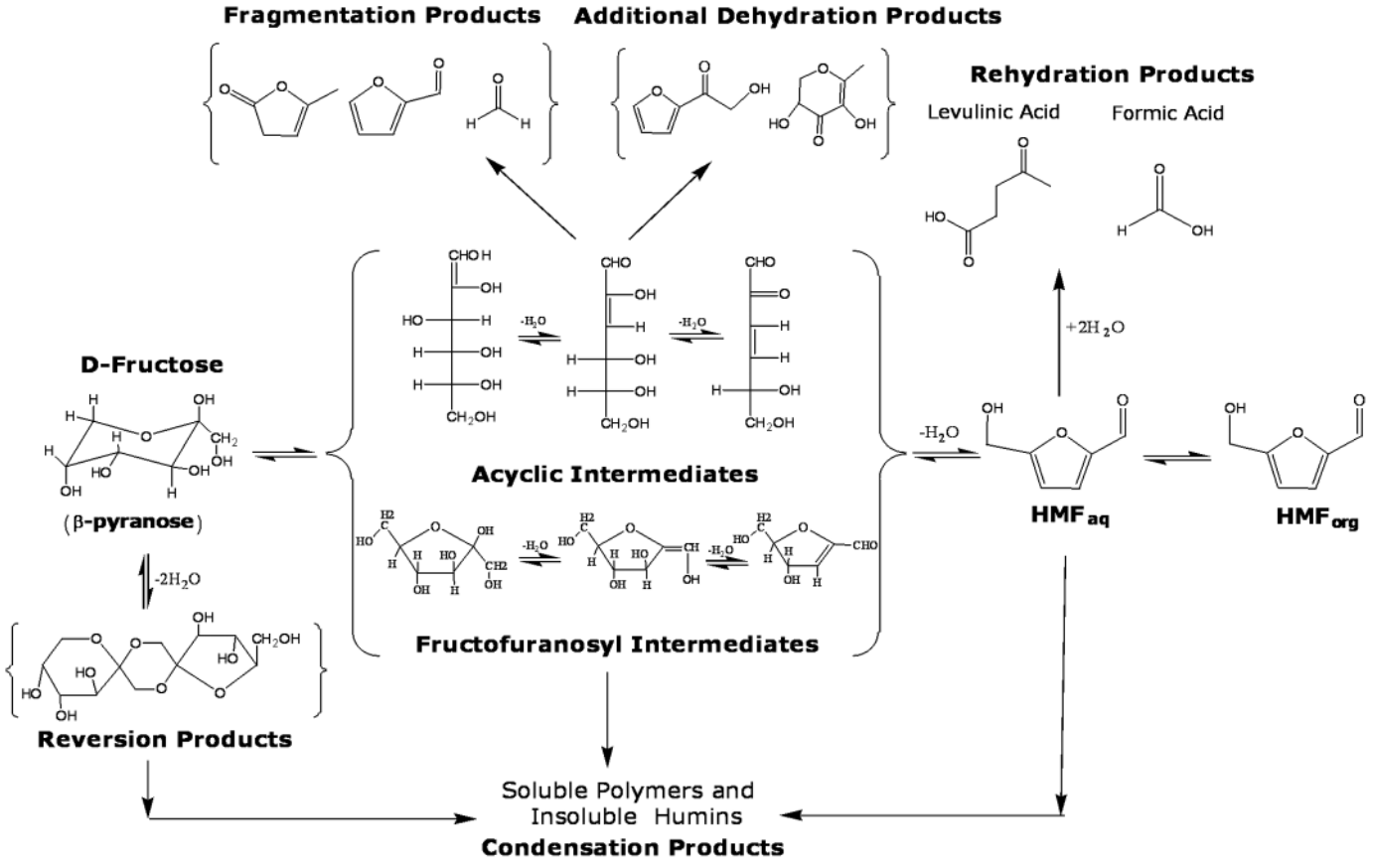



**Fig. 1.** Schematic representation of reaction pathways for acid-catalyzed dehydration of fructose. Structures in brackets correspond to representative species. Detailed chemistry is presented in (*20*, *23*).

intermediate is currently impeded by high production costs. Early work showed that HMF could be produced in high yields using high-boiling organic solvents [such as dimethylsulfoxide (DMSO), dimethylformamide, and mixtures of polyethyleneglycol with water] over various catalysts, including sulfuric acid and sulfonic acid resins. However, this approach necessitates difficult and energy-intensive isolation procedures (*6*, *9*–*13*). In pure water, fructose dehydration is generally nonselective, leading to many by-products besides HMF (*14*). Recent advances have shown improved results in pure water or in water-miscible solvent systems (such as acetonitrile or acetone), but only with the use of low initial fructose concentrations, which inevitably generate low HMF concentrations (*1*, *10*, *15*, *16*). Biphasic systems, in which a

**Table 1.** Results for acid-catalyzed dehydration of fructose. Runs 1 to 27 were carried out at 453 K for 2.5 to 3 min with 0.25 M HCl aqueous phase solutions; runs 28 to 39 were carried out at 363 K for 8 to 16 hours with an acidic ion-exchange resin at a 1:1 w/w fructose:resin ratio. Aqueous phase and organic phase compositions are reported as w/w ratios. Conversion is defined as the ratio of fructose consumed to fructose added initially. $R$ = $[HMF]_{org}/[HMF]_{aq}$. Standard runs for HCl, $H_2SO_4$, and $H_3PO_4$ catalysts used 1.5 g of aqueous phase and 1.5 g of extracting solvent. Runs marked with an asterisk used 3 g of extracting solvent. Runs for resin catalyst used 5.0 g of aqueous phase and 5.0 g of extracting solvent. $V_{org}/V_{aq}$ was measured upon completion of reaction.

| Run no. | Aqueous phase composition | Organic phase composition | Conversion (%) | Selectivity (%) | $[HMF]_{aq}$ (g/ml) | $[HMF]_{org}$ (g/ml) | $R$ | $V_{org}/V_{aq}$ |
|---|---|---|---|---|---|---|---|---|
| | | | *30 wt % fructose with HCl catalyst* | | | | | |
| 1 | Water | none | 50 | 51 | 0.060 | – | 0.00 | 0.00 |
| 2 | Water | MIBK | 91 | 60 | 0.056 | 0.050 | 0.90 | 1.51 |
| 3* | Water | MIBK | 75 | 73 | 0.035 | 0.033 | 0.96 | 3.13 |
| 4 | Water | 7:3 MIBK:2-butanol | 68 | 70 | 0.033 | 0.054 | 1.65 | 1.56 |
| 5* | Water | 7:3 MIBK:2-butanol | 86 | 80 | 0.026 | 0.045 | 1.73 | 3.68 |
| 6 | 8:2 Water:DMSO | MIBK | 94 | 67 | 0.077 | 0.050 | 0.66 | 1.41 |
| 7 | 8:2 Water:DMSO | 7:3 MIBK:2-butanol | 80 | 75 | 0.050 | 0.064 | 1.30 | 1.49 |
| 8* | 8:2 Water:DMSO | 7:3 MIBK:2-butanol | 87 | 82 | 0.034 | 0.046 | 1.39 | 3.65 |
| 9 | 7:3 Water:PVP | MIBK | 74 | 66 | 0.055 | 0.041 | 0.81 | 1.56 |
| 10 | 7:3 Water:PVP | 7:3 MIBK:2-butanol | 62 | 76 | 0.042 | 0.047 | 1.25 | 1.57 |
| 11* | 7:3 Water:PVP | 7:3 MIBK:2-butanol | 79 | 82 | 0.030 | 0.041 | 1.44 | 3.83 |
| 12 | 7:3(8:2 Water:DMSO):PVP | MIBK | 79 | 75 | 0.071 | 0.047 | 0.71 | 1.52 |
| 13 | 7:3(8:2 Water:DMSO):PVP | 7:3 MIBK:2-butanol | 82 | 83 | 0.063 | 0.065 | 1.12 | 1.62 |
| 14* | 7:3(8:2 Water:DMSO):PVP | 7:3 MIBK:2-butanol | 89 | 85 | 0.043 | 0.046 | 1.17 | 3.99 |
| | | | *50 wt % fructose with HCl catalyst* | | | | | |
| 15 | Water | none | 51 | 28 | 0.064 | – | 0.00 | 0.00 |
| 16 | Water | MIBK | 65 | 47 | 0.049 | 0.051 | 1.11 | 1.80 |
| 17 | Water | 7:3 MIBK:2-butanol | 71 | 59 | 0.049 | 0.079 | 1.73 | 1.91 |
| 18* | Water | 7:3 MIBK:2-butanol | 88 | 72 | 0.045 | 0.069 | 1.55 | 4.66 |
| 19 | 8:2 Water:DMSO | MIBK | 71 | 57 | 0.076 | 0.060 | 0.86 | 1.69 |
| 20 | 8:2 Water:DMSO | 7:3 MIBK:2-butanol | 80 | 63 | 0.077 | 0.085 | 1.19 | 1.87 |
| 21* | 8:2 Water:DMSO | 7:3 MIBK:2-butanol | 91 | 74 | 0.059 | 0.072 | 1.30 | 4.87 |
| 22 | 7:3 Water:PVP | MIBK | 85 | 56 | 0.074 | 0.060 | 0.80 | 1.72 |
| 23 | 7:3 Water:PVP | 7:3 MIBK:2-butanol | 77 | 61 | 0.076 | 0.081 | 1.19 | 1.85 |
| 24* | 7:3 Water:PVP | 7:3 MIBK:2-butanol | 90 | 77 | 0.062 | 0.070 | 1.22 | 5.15 |
| 25 | 7:3(8:2 Water:DMSO):PVP | MIBK | 77 | 61 | 0.095 | 0.066 | 0.77 | 1.85 |
| 26 | 7:3(8:2 Water:DMSO):PVP | 7:3 MIBK:2-butanol | 72 | 62 | 0.068 | 0.074 | 1.25 | 1.89 |
| 27* | 7:3(8:2 Water:DMSO):PVP | 7:3 MIBK:2-butanol | 92 | 77 | 0.076 | 0.070 | 1.03 | 5.11 |
| | | | *10 wt % fructose with ion-exchange resin catalyst* | | | | | |
| 28 | Water | MIBK | 75 | 44 | 0.010 | 0.011 | 1.02 | 1.32 |
| 29 | Water | MIBK | 17 | 43 | 0.0021 | 0.0024 | 1.15 | 1.29 |
| 30 | Water | 7:3 MIBK:2-butanol | 61 | 60 | 0.0086 | 0.014 | 1.61 | 1.31 |
| 31 | 8:2 Water:DMSO | MIBK | 84 | 47 | 0.015 | 0.012 | 0.79 | 1.26 |
| 32 | 8:2 Water:DMSO | MIBK | 19 | 80 | 0.0052 | 0.0045 | 0.87 | 1.24 |
| 33 | 8:2 Water:DMSO | 7:3 MIBK:2-butanol | 74 | 68 | 0.015 | 0.017 | 1.18 | 1.24 |
| 34 | 7:3 Water:PVP | MIBK | 74 | 63 | 0.018 | 0.013 | 0.79 | 1.43 |
| 35 | 7:3 Water:PVP | 7:3 MIBK:2-butanol | 70 | 65 | 0.015 | 0.015 | 1.04 | 1.46 |
| 36 | 7:3(8:2 Water:DMSO):PVP | MIBK | 80 | 71 | 0.026 | 0.013 | 0.54 | 1.38 |
| 37 | 7:3(8:2 Water:DMSO):PVP | 7:3 MIBK:2-butanol | 76 | 77 | 0.020 | 0.019 | 1.03 | 1.43 |
| | | | *30 wt % fructose with ion-exchange resin catalyst* | | | | | |
| 38 | 7:3(8:2 Water:DMSO):PVP | MIBK | 89 | 60 | 0.066 | 0.041 | 0.66 | 1.65 |
| 39 | 7:3(8:2 Water:DMSO):PVP | 7:3 MIBK:2-butanol | 83 | 65 | 0.053 | 0.051 | 1.07 | 1.74 |
| | | | *30 wt % fructose with $H_2SO_4$ catalyst* | | | | | |
| 40* | Water | 7:3 MIBK:2-butanol | 80 | 66 | 0.022 | 0.035 | 1.63 | 3.54 |
| 41* | 8:2 Water:DMSO | 7:3 MIBK:2-butanol | 85 | 71 | 0.029 | 0.040 | 1.35 | 3.59 |
| | | | *30 wt % fructose with $H_3PO_4$ catalyst* | | | | | |
| 42* | Water | 7:3 MIBK:2-butanol | 65 | 65 | 0.016 | 0.029 | 1.89 | 3.47 |
| 43* | 8:2 Water:DMSO | 7:3 MIBK:2-butanol | 51 | 76 | 0.016 | 0.025 | 1.58 | 2.95 |

water-immiscible organic solvent is added to extract continuously the HMF from the aqueous phase, have also been investigated with the use of mineral acid or zeolite catalysts at temperatures greater than 450 K (*6*, *17*–*21*). However, poor HMF partitioning into the organic streams used in these studies necessitated large amounts of solvent, thereby requiring large energy expenditures to purify the diluted HMF product (*22*).

Fructose is dehydrated in the presence of an acid catalyst to produce HMF and various by-products (Fig. 1). Although evidence exists supporting both the open-chain and the cyclic fructofuransyl intermediate pathways (*20*, *23*), it is clear that the reaction intermediates and the HMF product degrade by means of processes such as fragmentation, condensation, rehydration, reversion, and/or additional dehydration reactions. We used a modified biphasic system to address key HMF production limitations. Specifically, we improved the HMF selectivity (defined as the moles of HMF produced divided by the moles of fructose reacted) of the acid-catalyzed dehydration of concentrated [30 to 50 weight % (wt %)] fructose feeds by adding modifiers to both phases. The reactive aqueous phase containing the catalyst and the sugar was modified with polar aprotic molecules [DMSO or 1-methyl-2-pyrrolidinone (NMP)] and a hydrophilic polymer [poly(1-vinyl-2-pyrrolidinone) (PVP)]. The water-immiscible organic phase [methylisobutylketone (MIBK)] used during the reaction to extract HMF was modified with 2-butanol. The ratio of relative volumes of the organic and aqueous phases in the reactor ($V_{org}/V_{aq}$), as well as the ratio of the HMF concentration in the organic layer to that in the aqueous layer (defined as the extraction ratio, *R*) proved to be important variables in the process. Upon completion of the dehydration reaction, both phases can be separated for efficient product isolation. Although various acid catalysts can be used to perform the dehydration reaction, HCl showed the highest HMF selectivity of the common mineral acid catalysts (Table 1, runs 5, 8, and 40 to 43).

We performed experiments using 0.25 M HCl at 453 K under autonomous pressure. To optimize selectivity, we varied the concentrations of each aqueous phase modifier (Table 1 and Fig. 2) (*24*). Together, DMSO and PVP increased the selectivity from 60 to 75% (Fig. 2A). However, we also sought to optimize partitioning of the HMF product into the organic phase both to minimize degradation reactions arising from extended HMF residence in the reactive aqueous phase and to achieve more efficient recovery of HMF in the subsequent isolation step. Unfortunately, the aqueous additives are a liability in this respect, because they increase the solubility of HMF in the aqueous phase (thereby decreasing the *R* value). The addition of 2-butanol to the organic phase helped counteract this effect by increasing HMF solubility in the organic phase relative to pure MIBK (Fig. 2B). Although pure 2-butanol is moderately soluble in water [up to 19.5 wt % at 298 K, as measured by high-performance liquid chromatography (HPLC)], it partitioned effectively into the MIBK layer when mixed with MIBK at a 3:7 weight/weight (w/w) ratio (<5 wt % at 298 K was detected in the aqueous phase by HPLC). Starting with a 30 wt % aqueous fructose solution, our optimal results using all three modifiers (DMSO, PVP, and 2-butanol) yielded 0.065 g/ml of HMF in the organic layer, with 83% HMF selectivity at 82% conversion (Table 1, run 13).

The selectivity for production of HMF decreases when the initial sugar concentration is increased from 30 to 50 wt % (Fig. 2C). This result is in agreement with literature reports that increasing the fructose concentration leads to higher rates of condensation reactions (*6*, *13*). The addition of DMSO, PVP, and/or 2-butanol to the 50 wt % fructose systems in the same quantities used for lower fructose concentrations did not substantially improve the HMF selectivity. However, doubling the amount of the 7:3 MIBK:2-butanol extracting solvent increased the selectivity substantially. The optimal system using all three modifiers generated 0.070 g/ml of HMF in the organic phase at 77% HMF selectivity and 92% conversion (Table 1, run 27).

A heterogeneous catalyst is more easily separated from the product and recycled than a homogeneous catalyst, leading us to investigate fructose dehydration with the use of an acidic ion-exchange resin. These experiments were conducted at 363 K under autonomous pressure (Table 1, runs 28 to 39). When we used the resin catalyst, we observed trends in modifier impact that were similar to those we saw with HCl; however, the absolute selectivities were lower for the resin system. Kuster (*6*) reported that elevated temperatures (>453 K) favor higher HMF selectivity because at these conditions the rate of HMF formation is 4 to 10 times as fast as the rate of HMF degradation. Because ion-exchange resins degrade at these higher temperatures, it is desirable to replace HCl with a heterogeneous acid catalyst that is stable to 450 K. In this respect, and considering the favorable results we obtained using $H_3PO_4$ (Table 1, run 43), we tested a niobium phosphate catalyst at 453 K, obtaining promising results of 73% HMF selectivity at 62% conversion. In addition, Dias *et al*. (*25*) have shown that nanoporous materials possessing sulfonic acid groups can promote dehydration of xylose to furfural at elevated temperatures. Carlini *et al*. (*26*) observed fructose dehydration at low temperatures (353 K) and mostly at low conversion (<50%) with the use of vanadyl-phosphate–based catalysts, which may also withstand elevated temperatures.

Experiments conducted at low temperature and low conversions (Table 1, runs 29

**Table 2.** Simulation of HMF yield (*Y*) and energetic yield ($Y\eta$) for selected dehydration systems. $[HMF]_{aq}$ corresponds to the HMF concentration in the aqueous phase leaving the extractor, and $[HMF]_{org}$ corresponds to the HMF concentration entering the evaporator in Fig. 3. Runs are based on runs in Table 1. The selectivity is set to the value obtained experimentally, and conversion is assumed to be 90%. The yield is calculated based on the HMF present in the organic stream sent to the evaporator.

| Run no. | Aqueous phase composition | Organic phase composition | Selectivity (%) | $[HMF]_{aq}$ (g/ml) | $[HMF]_{org}$ (g/ml) | *Y* (%) | $Y\eta$ (%) |
|---|---|---|---|---|---|---|---|
| | | | *30 wt % fructose* | | | | |
| 2 | Water | MIBK | 60 | 0.0075 | 0.045 | 48 | 34 |
| 4 | Water | 7:3 MIBK:2-butanol | 70 | 0.0001 | 0.057 | 61 | 43 |
| 6 | 8:2 Water:DMSO | MIBK | 67 | 0.025 | 0.048 | 48 | 35 |
| 7 | 8:2 Water:DMSO | 7:3 MIBK:2-butanol | 75 | 0.0009 | 0.063 | 66 | 48 |
| 12 | 7:3(8:2 Water:DMSO):PVP | MIBK | 75 | 0.024 | 0.057 | 56 | 44 |
| 13 | 7:3(8:2 Water:DMSO):PVP | 7:3 MIBK:2-butanol | 83 | 0.0032 | 0.071 | 73 | 56 |
| | | | *50 wt % fructose* | | | | |
| 16 | Water | MIBK | 47 | 0.0026 | 0.054 | 39 | 27 |
| 26 | 7:3(8:2 Water:DMSO):PVP | 7:3 MIBK:2-butanol | 62 | 0.0019 | 0.091 | 53 | 43 |
| 27 | 7:3(8:2 Water:DMSO):PVP | 7:3 MIBK:2-butanol | 77 | 0.0055 | 0.071 | 67 | 51 |

and 32) offer mechanistic insight about the role of DMSO in enhancing HMF selectivity. The results show that DMSO primarily increases the rate of fructose conversion into HMF and to some extent decreases the rates of undesirable parallel reactions. Likewise, earlier work has suggested that DMSO suppresses both the formation of condensation by-products and HMF rehydration by lowering the overall water concentration (*6*, *13*). Similar studies have indicated that DMSO both favorably shifts the equilibrium of the rate-determining step in HMF production and inhibits acyclic reaction sequences that may lead to undesirable intermediates (*23*).

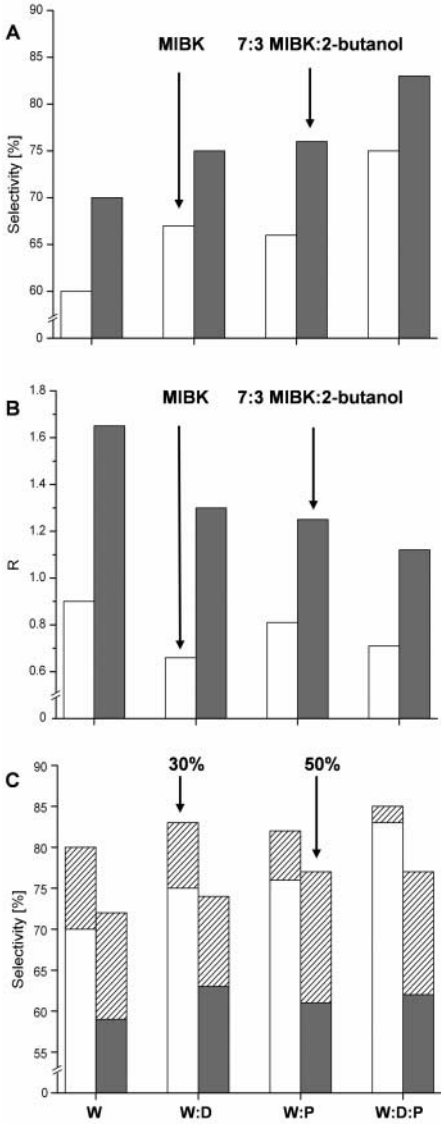



**Fig. 2.** Effects of changing the aqueous phase composition from water (W), to 8:2 w/w water:DMSO (W:D), to 7:3 w/w water:PVP (W:P), to 7:3 w/w (8:2 water:DMSO):PVP (W:D:P). (**A**) HMF selectivity with 30 wt % fructose feed. White and gray bars represent MIBK and MIBK:2-butanol extracting solvents, respectively. (**B**) Extraction ratio, *R*. White and gray bars represent MIBK and MIBK:2-butanol extracting solvents, respectively. (**C**) HMF selectivity with MIBK:2-butanol extracting solvent. White and gray bars represent 30 and 50 wt % fructose, respectively, and hatched bars show the improvement obtained with the use of double the amount of extracting solvent.

The reaction of 10 wt % fructose in 7:3 water:NMP with MIBK as the extracting solvent and an acidic ion-exchange resin catalyst generated 68% HMF selectivity at 80% conversion, indicating that the addition of NMP to water enhances the selectivity of HMF production from fructose. Because NMP is an aprotic solvent with properties similar to DMSO, NMP appears to act through mechanisms similar to those of DMSO to enhance HMF selectivity in the fructose dehydration reaction. However, whereas the carryover of DMSO from the aqueous phase into the organic phase is not considerable (<0.8 wt % DMSO in MIBK after contacting an 8:2 water:DMSO aqueous solution, as measured by HPLC), the carryover of NMP into the organic phase is high (~5 wt % NMP in MIBK after contacting a 7:3 water:NMP aqueous solution, as measured by HPLC), thereby complicating the subsequent separation of HMF from the organic phase by evaporation. Notably, we found that replacing NMP with PVP, a stable hydrophilic polymer that has NMP moieties along the polyethylene chain, preserves the benefits on selectivity produced by NMP but eliminates organic phase contamination due to the low solubility of PVP in the extracting solvent. In an analogous manner, grafting DMSO onto a hydrophilic polymeric backbone could be used as a strategy to eliminate trace amounts of DMSO in the organic phase.

The addition of 2-butanol to MIBK enhances HMF selectivity by removing HMF more efficiently from the reactive aqueous medium. Notably, increasing the extraction ratio *R* and/or increasing $V_{org}/V_{aq}$ can counteract the faster rate of HMF degradation in the presence of fructose. This undesirable reaction between fructose and HMF is reflected in lower HMF selectivities at 50 wt % fructose when compared with 30 wt % (Table 1), and we observed directly that lower selectivities are obtained when controlled amounts of HMF are added initially to the fructose reaction system. In addition, separating HMF from the aqueous medium lowers the rate of HMF rehydration into levulinic and formic acids. Analyses by gas chromatography–mass spectrometry (GC-MS) of the aqueous and organic phases after conversion of 30 wt % fructose showed that the general composition of the by-products corresponds to typically 10% rehydration, 5% dehydration, 5% fragmentation, and 80% condensation compounds.

We performed simulations for selected experiments from Table 1 to estimate the HMF concentrations that would be obtained by combining our batch reactor experiments with a countercurrent extractor to remove the HMF remaining in the aqueous layer (Fig. 3). The final amount of HMF obtained by combining the organic streams from the reactor and the extractor (i.e., the stream entering the evaporator) is used to calculate the energetic yield ($Y\eta$) as a measure of the overall efficiency of our process for obtaining

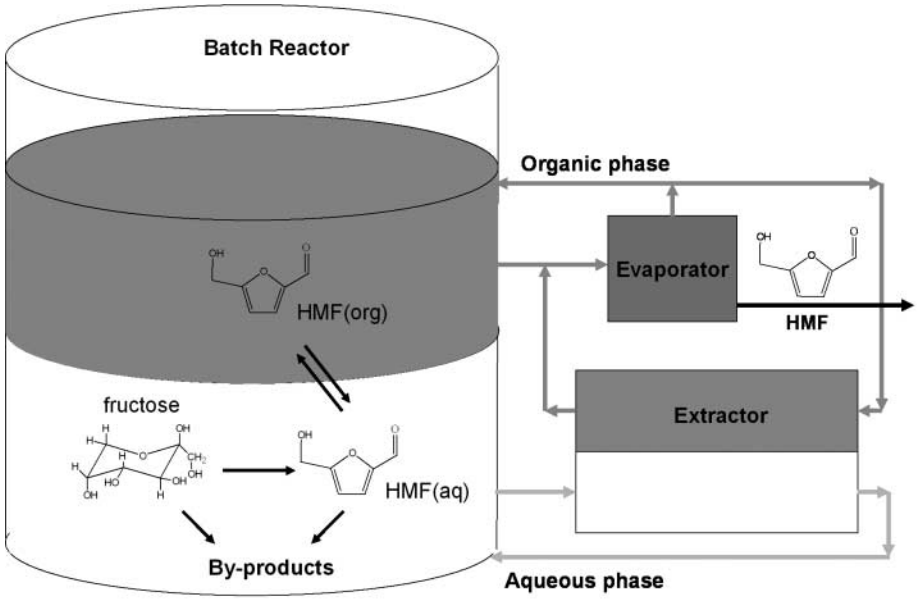



**Fig. 3.** Batch process for production of HMF from fructose with simulated countercurrent extraction and evaporation steps. The aqueous phase (white) contains fructose, DMSO, PVP, and the acid catalyst and is represented in the bottom half of the batch reactor. The organic phase (gray) contains MIBK or MIBK:2-butanol and is represented in the top half of the batch reactor.

HMF by solvent evaporation. The energetic yield is the product of the HMF yield ($Y$), defined as the moles of HMF in the stream entering the evaporator in Fig. 3 divided by the total moles of fructose fed to the batch reactor, and an energy efficiency ($\eta$), defined as the heat of combustion of the HMF product ($\Delta H_{C,HMF}$) minus the energy necessary to evaporate the solvent ($\Delta H_{vap,org}$), normalized by the energy content of the product [i.e., $\eta = (\Delta H_{C,HMF} - \Delta H_{vap,org})/\Delta H_{C,HMF}$]. These simulations used the experimental selectivity for each system (Table 1), which was assumed to remain constant at 90% conversion, the experimental value of $V_{org}/V_{aq}$ for the batch reactor, and the experimental value of $R$ to model a countercurrent extractor operating with equal volumes of aqueous and organic streams. Aqueous and organic phase modifiers improve the value of $Y\eta$, thus reducing energy expenditures required to obtain the HMF product when compared with the water/MIBK system (Table 2).

The value of $Y\eta$ alone does not address the difficulties of using high-boiling organic systems. For example, although a theoretical value of $Y\eta > 75\%$ can be obtained with pure DMSO, the HMF product cannot be separated from the solvent by simple evaporation. Previous work has shown that because of the reactive nature of concentrated HMF at high temperatures, distillation of HMF from DMSO leads to substantial carbonization of the product (*10*). Low-temperature separation processes such as vacuum evaporation and vacuum distillation have been used to separate various solvents and by-products from HMF mixtures, but no experimental data have been reported for DMSO (*27*–*29*). Accordingly, we used Aspen Plus Simulation Software (Version 12.1, Aspen Technology Inc.) to compare energy requirements for the separations of HMF from a low-boiling solvent (pure MIBK) and from a high-boiling solvent (pure DMSO) for vacuum evaporation and vacuum distillation processes (for HMF levels of 10 wt %). Vacuum evaporation simulations predicted that 99.5% of the MIBK solvent can be evaporated at 13 mbar and 343 K with a 2.5% loss of HMF, whereas evaporating DMSO at 1.3 mbar and the same temperature resulted in a 30% loss of HMF. Consequently, HMF separation from DMSO with minimal losses requires the more expensive vacuum distillation process (e.g., 0.66 mbar and a bottom temperature of 386 K). When comparing both solvents with the use of vacuum distillation, simulations predicted that an efficient separation of HMF from pure DMSO requires 40% more energy as compared with pure MIBK, clearly showing the advantages of using a low-boiling solvent system.

### References and Notes


1. M. Bicker, J. Hirth, H. Vogel, *Green Chem.* **5**, 280 (2003).
2. T. Werpy, G. Petersen, "Top Value Added Chemicals From Biomass," *Technical Report No. DOE/GO-102004-1992* (National Renewable Energy Lab, Golden, CO, 2004) (available at www.osti.gov/bridge).
3. K. W. Pentz, British Patent 2, 131, 014 (1984).
4. G. W. Huber, J. N. Chheda, C. J. Barrett, J. A. Dumesic, *Science* **308**, 1446 (2005).
5. C. Moreau, M. N. Belgacem, A. Gandini, *Top. Catal.* **27**, 11 (2004).
6. B. M. F. Kuster, *Starch* **42**, 314 (1990).
7. A. Gaset, J. P. Gorrichon, E. Truchot, *Inf. Chim.* **212**, 179 (1981).
8. J. Lewkowski, *Arkivoc* **1**, 17 (2001) (available at www.arkat-usa.org/home.aspx?VIEW=MANUSCRIPT&MSID=403).
9. Y. Nakamura, S. Morikawa, *Bull. Chem. Soc. Jpn.* **53**, 3705 (1980).
10. D. W. Brown, A. J. Floyd, R. G. Kinsman, Y. Roshan-Ali, *J. Chem. Technol. Biotechnol.* **32**, 920 (1982).
11. H. H. Szmant, D. D. Chundury, *J. Chem. Technol. Biotechnol.* **31**, 135 (1981).
12. K. Seri, Y. Inoue, H. Ishida, *Bull. Chem. Soc. Jpn.* **74**, 1145 (2001).
13. H. E. van Dam, A. P. G. Kieboom, H. van Bekkum, *Starch* **38**, 95 (1986).
14. K. M. Rapp, U.S. Patent 4, 740, 605 (1987).
15. F. Benvenuti *et al.*, *Appl. Catal. A Gen.* **193**, 147 (2000).
16. C. Carlini *et al.*, *Appl. Catal. A Gen.* **183**, 295 (1999).
17. Q. P. Peniston, U.S. Patent 2, 750, 394 (1956).
18. T. El Hajj, A. MasRoua, J. C. Martin, G. Descotes, *Bull. Soc. Chim. Fr.* **5**, 855 (1987).
19. L. Rigal, A. Gaset, J.-P. Gorrichon, *Ind. Eng. Chem. Prod. Res. Dev.* **20**, 719 (1981).
20. C. Moreau *et al.*, *Appl. Catal. A Gen.* **145**, 211 (1996).
21. P. Rivalier, J. Duhamet, C. Moreau, R. Durand, *Catal. Today* **24**, 165 (1995).
22. The HMF yield and HMF concentration (in units of g/ml) as reported by different authors in representative systems are presented as follows: Dehydration system: HMF yield, HMF concentration (reference numbers). DMSO: >95%, <0.13 (*9*–*12*); polyethylene glycol/water: 60%, 0.28 (*6*, *13*); water, 34%, <0.06 (*14*); water and water-miscible solvents: >75%, <0.04 (*1*, *10*, *15*, *16*); biphasic systems: >75%, <0.02 (*6*, *17*–*20*).
23. M. J. J. Antal, W. S. L. Mok, G. N. Richards, *Carbohydr. Res.* **199**, 91 (1990).
24. Materials and methods are available as supporting material on *Science* Online.
25. A. S. Dias, M. Pillinger, A. A. Valente, *J. Catal.* **229**, 414 (2005).
26. C. Carlini, P. Patrono, A. M. R. Galletti, G. Sbrana, *Appl. Catal. A Gen.* **275**, 111 (2004).
27. J. F. Harris, J. F. Saeman, L. L. Zoch, *Forest Prod. J.* **10**, 125 (1960).
28. R. H. Hunter, U.S. Patent 3, 201, 331 (1965).
29. R. E. Jones, H. B. Lange, U.S. Patent 2, 994, 645 (1958).


30. This work was supported by the U.S. Department of Agriculture and the NSF Chemical and Transport Systems Division of the Directorate for Engineering.


### Supporting Online Material

www.sciencemag.org/cgi/content/full/312/5782/1933/DC1
Materials and Methods

16 February 2006; accepted 30 May 2006
10.1126/science.1126337

# An Octahedral Coordination Complex of Iron(VI)

**John F. Berry,[1]* Eckhard Bill,[1] Eberhard Bothe,[1] Serena DeBeer George,[2] Bernd Mienert,[1] Frank Neese,[1]† Karl Wieghardt[1]‡**

**The hexavalent state, considered to be the highest oxidation level accessible for iron, has previously been found only in the tetrahedral ferrate dianion, $FeO_4^{2-}$. We report the photochemical synthesis of another Fe(VI) compound, an octahedrally coordinated dication bearing a terminal nitrido ligand. Mössbauer and x-ray absorption spectra, supported by density functional theory, are consistent with the octahedral structure having an Fe≡N triple bond of 1.57 angstroms and a singlet $d_{xy}^2$ ground electronic configuration. The compound is stable at 77 kelvin and yields a high-spin Fe(III) species upon warming.**

Iron, the most abundant transition element in the natural world, generally occurs in compounds of its divalent or trivalent ions. Iron ions in these two oxidation states occur in a wide range of minerals (*1*, *2*) found terrestrially as well as on Mars (*3*, *4*) and are widely used in biological electron transfer processes, such as those involving iron-sulfur clusters (*5*). Low-valent, electron-rich compounds play key roles in biological hydrogen production, for example, Fe(I) and Fe(0) (*6*–*8*). Nature also uses more electron-deficient iron centers for highly specific and efficient enzymatic oxidation (*9*–*11*), which has spurred chemists to prepare analogous coordination complexes for use in synthesis and oxidative waste remediation. Notable among these are Fe(IV) complexes that activate $H_2O_2$ (*12*) and Fe(IV)=O species that hydroxylate hydrocarbons (*13*, *14*), as well as several other Fe(IV) (*15*–*19*) and Fe(V) (*20*, *21*) compounds.

Although iron has eight electrons in its valence shell, only the +6 state is considered

[1]Max-Planck-Institut für Bioanorganische Chemie, Stiftstrasse 34-36, D-45470 Mülheim an der Ruhr, Germany. [2]Stanford Synchrotron Radiation Laboratory, Stanford University, Stanford, CA 94309, USA.

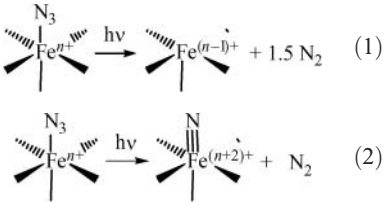
*Present address: Department of Chemistry, University of Wisconsin, 1101 University Avenue, Madison, WI 53706, USA.

†Present address: Institut für Physikalische und Theoretische Chemie, Universität Bonn, D-53115 Bonn, Germany.

‡To whom correspondence should be addressed. E-mail: wieghardt@mpi-muelheim.mpg.de

accessible and only a single Fe(VI) species is known. This ferrate ion, $[Fe^{VI}O_4]^{2-}$, is an especially powerful oxidant used in organic synthesis (*22*), soil and wastewater treatment (*23*, *24*), batteries (*25*), and disinfectants (*26*). Its reduction product, rust, is innocuous—an advantage over many other oxidants—but ferrate is unstable and often indiscriminately reactive. The search for additional hexavalent iron species that may react in a more specific and controlled way has been long and fruitless. Here, we report the low-temperature synthesis and characterization of a second Fe(VI) compound.

One successful approach to the synthesis of high-valent iron species has been photolysis of lower-valent azido complexes to yield either a photoreduced or photo-oxidized product:




Homolytic Fe–N bond cleavage in the reductive pathway (Eq. 1) returns one electron to the metal center and produces an azide radical that quickly decomposes to dinitrogen. In the oxidative pathway (Eq. 2), dinitrogen is formed via heterolytic N–N bond cleavage, leaving a terminal nitrido group and formally oxidizing the iron center by two electrons. This synthetic approach was first applied to iron-porphyrin species (*27*) and later to nonheme iron complexes (*20*, *21*). In our work, the Fe(III)-azido complex $[(cy\text{-}ac)FeN_3]PF_6$, **1**, where cy-ac is the anion 1,4,8,11-tetraazacyclotetradecane-1-acetate (Scheme 1), was photolyzed at 77 K with 420-nm light to produce the Fe(V)-nitrido species $[(cy\text{-}ac)FeN]PF_6$, **2** (*21*), which has been shown to have a nearly orbitally degenerate doublet ground state due to the Fe–N bond configuration $\sigma^2\pi^4\pi^{*1}$ (bond order 2.5) (*28*).

Recently, we reported the Fe(III)-azido complex **3** in which the N–H groups in the cy-ac ligand of **1** have been replaced by N–Me (methyl) groups (*29*). This change has important steric and electronic ramifications for the iron center, one of which is that **1** has a low-spin electronic ground state whereas **3** is high-spin (*30*, *31*). Another major difference between **1** and **3** is that **3** may be reversibly oxidized to the stable Fe(IV)-azido complex $[(Me_3cy\text{-}ac)FeN_3](PF_6)_2$, **4**, whereas electrochemical oxidation of **1** does not lead to a stable Fe(IV) product (*29*).

Although photolysis of **1** leads efficiently to the Fe(V) complex **2**, photolysis of **3** under identical conditions yields only a high-spin Fe(II) product, identified by Mössbauer spectroscopy. The isomer shift ($\delta$ = 1.04 mm $s^{-1}$) and quadrupole splitting ($\Delta E_Q$ = 2.59 mm $s^{-1}$) of the photoreduced complex suggest that the product is the octahedral species $[(Me_3cy\text{-}ac)Fe(NCCH_3)]^+$. No conditions have yet been found to induce photo-oxidation of **3**.

Compound **4** is also photochemically reactive. Upon laser irradiation at 77 K with 650-nm light, frozen solutions of **4** change color from cherry-red to pale yellow. A frozen Mössbauer sample of **4** (generated electrochemically in acetonitrile solution containing 0.1 M $NBu_4PF_6$, where Bu = butyl) was irradiated for 5 hours under these conditions, and the Mössbauer spectrum (Fig. 1A) is well simulated by two subspectra, neither of which corresponds to **4** ($\delta$ = 0.11 mm $s^{-1}$, $\Delta E_Q$ = 1.92 mm $s^{-1}$). Subspectrum b, which accounts for 27% of the iron in the sample, is broad, with $\delta$ = 0.53 mm $s^{-1}$ and $\Delta E_Q$ = 1.13 mm $s^{-1}$, consistent with a high-spin Fe(III) species that may form via photoreduction. The major species present (subspectrum a) has a very low $\delta$ value of –0.29 mm $s^{-1}$ and $\Delta E_Q$ of 1.53 mm $s^{-1}$. This $\delta$ value is 0.40 mm $s^{-1}$ lower than that of **4** and 0.19

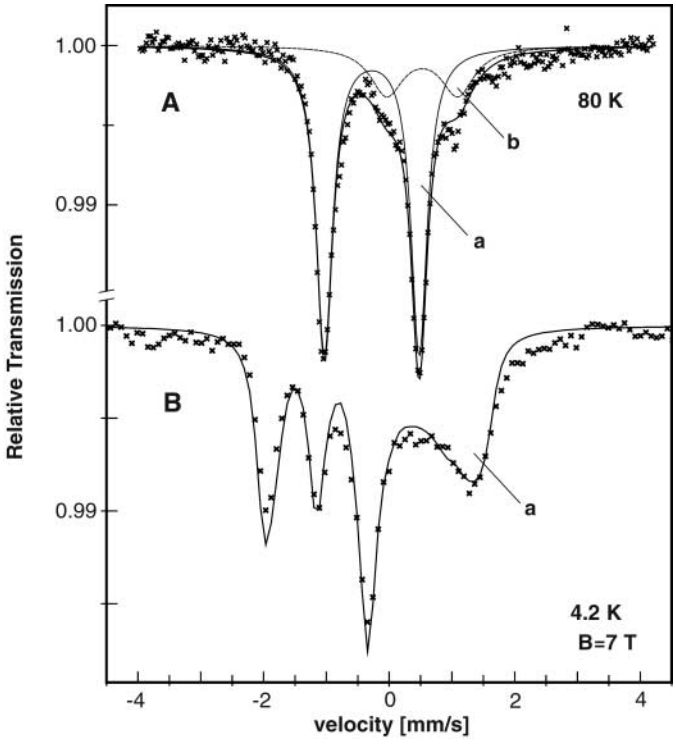
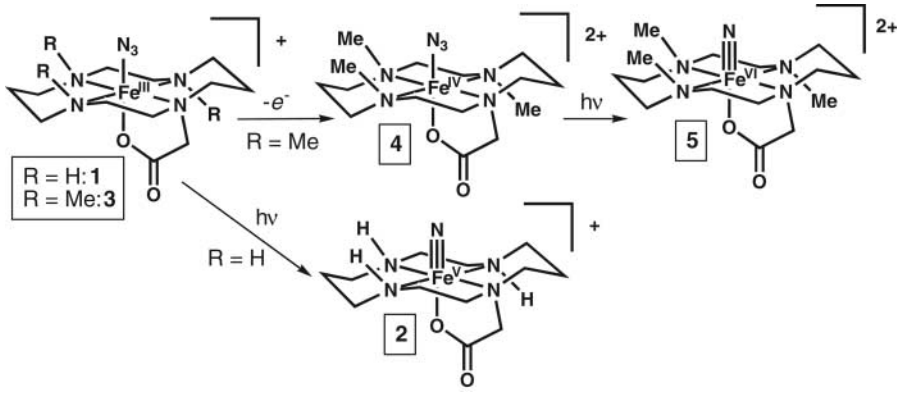



**Fig. 1.** Mössbauer spectra of a photolyzed acetonitrile solution of **4** measured at 80 K (**A**) and at 4.2 K in an applied magnetic field of 7 T (**B**). Subspectrum a corresponds to **5** and was the only subspectrum that was fitted in (B).




**Scheme 1.**

**Table 1.** Comparison of experimental and calculated properties of **5**. Errors are SDs, $\nu$(C=O) is the C=O stretching frequency, and $\eta$ represents the asymmetry parameter of the electric field gradient tensor.

| | Fe≡N (Å) | Fe–$N_{av}$ (Å) | Fe–O (Å) | $\nu$(C=O) ($cm^{-1}$) | $\delta$ (mm $s^{-1}$) | $\Delta E_Q$ (mm $s^{-1}$); $\eta$ |
|---|---|---|---|---|---|---|
| Experiment | 1.57 ± 0.02 | 2.03 ± 0.02* | 2.03 ± 0.02* | 1706 | –0.29 | +1.53; 0.33 |
| Calculation | 1.532 | 2.124 | 1.911 | 1740 | –0.31 | +0.75; 0.52 |

*Only an average Fe–N/O distance could be refined from the EXAFS data. The values 2.06 Å for the Fe–N bonds and 1.92 Å for the Fe–O bond are in reasonable agreement with this average value.

mm $s^{-1}$ lower than that of **2**, which suggests that species "a" has a higher formal oxidation state than either **4** or **2**. We therefore assign this species as $[(Me_3cy\text{-}ac)FeN](PF_6)_2$, **5**, an Fe(VI)-nitrido complex.

The δ value of **5** is considerably higher than that of the ferrate ion in various salts (–0.79 to –0.85 mm $s^{-1}$ at 80 K) (*32*), which is the only other Fe(VI) species known. Note that there are major differences between the geometric and electronic structures of these two species that prohibit the direct use of their δ values to infer electron deficiency. The most readily apparent difference is in the coordination number of the iron. The influence of the coordination number of the iron on the isomer shift is well demonstrated by comparing salts of the $[FeCl_6]^{3-}$ trianion with those of the $[FeCl_4]^-$ anion; the isomer shifts in the former cases are typically larger by 0.23 mm $s^{-1}$ (*33*), which reflects the difference in overall covalency due to different metal-ligand bond lengths. Also important is the fact that the ferrate dianion is paramagnetic with a spin triplet ground state ($S = 1$), but **5** is diamagnetic, as evidenced by its Mössbauer spectrum measured in an applied magnetic field of 7 T (Fig. 1B). Although the spin state often has a striking effect on the Mössbauer isomer shift in Fe(II) and Fe(III) species because of the population of σ antibonding orbitals in the high-spin states, typically this causes the high-spin species to have higher δ values due to longer metal-ligand bond lengths and lower covalency. In the case of $d^2$ Fe(VI) ions, we expect this effect to be diminished because the σ antibonding orbitals will not be populated in the high-spin state. Another major difference between ferrate and **5** is that there are four strong and covalent Fe=O double bonds in the ferrate ion, whereas **5** has only one strong iron-nitrogen multiple bond (see below), and this effect is likely the dominating influence in the variation of the isomer shifts of these species.

It is useful to compare the isomer shift of **5** with those of other low-spin octahedral complexes with similar supporting ligands (Fig. 2) and electronic configurations ranging from $d^6$ to $d^2$. Because the compounds within this series use similar supporting ligands and their valence electrons exclusively occupy the $t_{2g}$-based orbitals, a nearly linear correlation between the isomer shift and oxidation state is observed, which further supports the assignment of **5** as a genuine Fe(VI) complex.

Electrochemical oxidation of **3** to **4** and subsequent photolysis to **5** was also monitored by infrared spectroscopy using an optically transparent thin layer electrode (OTTLE) cell. The spectral changes from **3** to **4** were as described in (*29*). Upon irradiation with laser light at 650 nm, a single new species is observed on the basis of a C=O stretching band at 1706 $cm^{-1}$ that grows concurrently with the diminishing intensity of the $N_3$ stretching band of **4** (*34*). We have determined a linear correlation between the C=O stretching frequencies of $Me_3$cy-ac complexes and the Fe-O bond distance (*35*), and the C=O stretch in **5** correlates well with an Fe–O bond distance of ~1.9 Å (fig. S1), which is fairly long for a bond to a highly oxidized metal center but is consistent with a bond trans to a short and strong iron-nitrogen multiple bond.

To further confirm our assignment of **5** as an Fe(VI)-nitrido complex, we analyzed the x-ray absorption spectrum. The edge and pre-edge region are shown to the left in Fig. 3, where they are compared to the corresponding data for the Fe(V) complex **2** (*28*). There is an intense pre-edge peak in each case, consistent with the presence of a strongly covalent iron-ligand multiple bond, and the area of this peak for **5** (39 ± 2 units) is significantly higher than for **2** (27 ± 2 units) (*28*). The peak area is also higher than those observed for Fe(IV)-oxo species (*36*, *37*), reflecting a greater covalency of the Fe≡N bond in **5**. The energy of this peak in **5** (7115.1 eV), as well as of the edge itself (7125.0 eV), is ~1 eV higher in energy than it is for **2** (7114.2 eV for the pre-edge and 7124.1 eV for the edge), consistent with an oxidation state one unit higher in the former. The pre-edge energies of **1**, **2**, **5**, and of the Fe(IV)-oxo complexes reported by Que (*37*) [averaged and recalibrated as noted in (*28*)] show a nearly linear relationship with oxidation state (Fig. 2).

The extended x-ray absorption fine structure (EXAFS) region of **5** was also analyzed

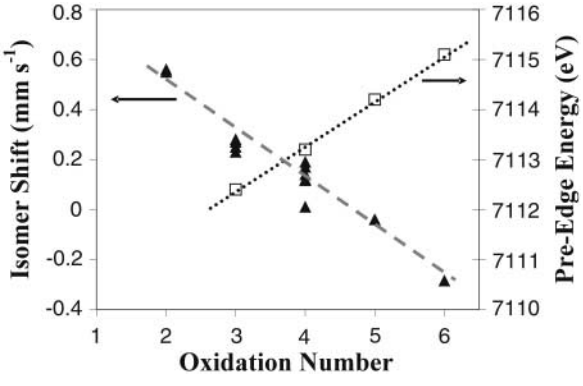



**Fig. 2.** Plot of the isomer shifts versus Fe oxidation state for the following low-spin iron complexes with cyclam-related ligands: *trans*-(cyclam)$Fe(N_3)_2$ (*20*), (cy-ac)$FeN_3$ (*21*), *trans*-[(cyclam)$Fe(N_3)_2$]$^+$ (*20*), [(cy-ac)FeX]$^+$ with X = $N_3$, Cl, OTf ($CF_3SO_3$), and F (*21*, *30*), low-spin $[(Me_3cy\text{-}ac)FeN_3]^+$ (*29*), [*trans*-(cyclam)$FeN_3]_2N$ (*20*), $[(cy\text{-}ac)Fe]_2N$ (*21*), $[(Me_4cyclam)FeO]^{2+}$ (*43*), $[(cy\text{-}ac)FeO]^+$ (*21*), $[(TMCS)FeO]^+$ (TMCS = 1-mercaptoethyl-4,8,11-trimethyl-1,4,8,11-tetraazacyclotetradecane) (*36*), $[(Me_3cy\text{-}ac)FeN_3]^{2+}$ (*29*), [*trans*-(cyclam)$FeN(N_3)$]$^+$ (*20*), $[(cy\text{-}ac)FeN]^+$ (*21*), and **5**. A plot of the Fe x-ray absorption spectroscopy (XAS) pre-edge peak energies of Fe(III), Fe(IV), Fe(V), and Fe(VI) complexes versus Fe oxidation state is also shown (white squares). The dashed and dotted lines represent least-squares fits [the Fe(IV) complex [(cy-ac)FeO]OTf is excluded in the former].

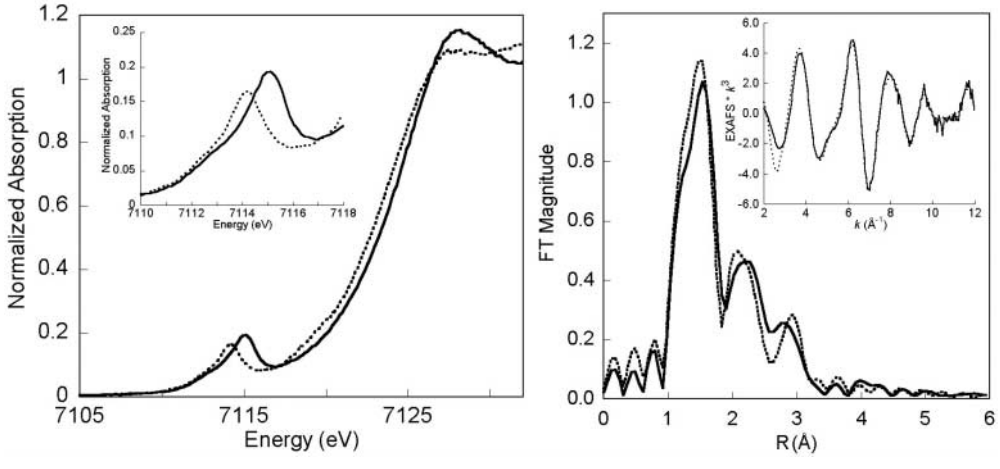



**Fig. 3.** (**Left**) Normalized Fe K-edge XAS spectra of **2** [dotted line, data taken from (*28*)] and **5** (solid line) with an expansion of the pre-edge region (inset). (**Right**) Non–phase-shift–corrected Fourier transform of the EXAFS data of **5** (inset). The solid lines are the data; dotted lines represent fits of the data.

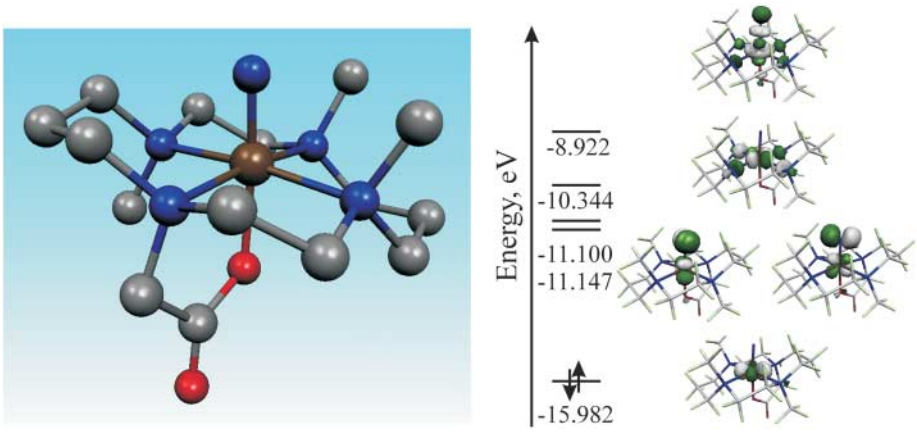



**Fig. 4.** (**Left**) Structure of the dication of **5** from DFT calculations with hydrogen atoms omitted. Color code: brown, iron; gray, carbon; blue, nitrogen; red, oxygen. (**Right**) Calculated 3d orbital manifold.

(Fig. 3, right, and table S1) and is best fit with a model in which there is one short Fe–N bond length of 1.57 Å and an average of five Fe–N or Fe–O bond distances of 2.03 Å [which could not be deconvoluted into separate Fe–N and Fe–O distances (*38*)]. The spectrum is consistent with the overall geometry predicted by a density functional theory (DFT) calculation on the dication $[(Me_3cy\text{-}ac)FeN]^{2+}$ (Fig. 4, left, and Table 1).

The calculated Fe≡N distance of 1.53 Å agrees well with the EXAFS result, and the Fe–O distance of 1.90 Å, which is elongated as a result of the trans influence of the short iron-nitrogen triple bond, agrees well with the infrared data and the absence of acetate-to-iron charge transfer bands in the visible spectrum of **5**. These electronic absorptions are observed in the visible spectrum of **4**, which appears to have a shorter Fe–O distance. The calculated Fe–N distances are elongated by ~0.06 to 0.09 Å relative to the EXAFS data (*38*), but this discrepancy is a systematic error in the calculation that has been observed in the calculated geometries of several other $Me_3$cy-ac complexes (*30*).

Vibrational and Mössbauer properties were also calculated for the dication $[(Me_3cy\text{-}ac)FeN]^{2+}$ (Table 1) and agree well with the values found experimentally. The calculated C=O stretch is calculated to be ~30 $cm^{-1}$ higher in energy than the observed value, which is a systematic overestimation for $Me_3$cy-ac complexes (fig. S1). Although an experimental value is lacking, the computed Fe≡N stretching frequency of 1064 $cm^{-1}$ in **5** is similar to that observed in the isoelectronic complex $[(cy\text{-}ac)Mn^{V}N]PF_6$, 1011 $cm^{-1}$ (*39*), and also in the Fe(IV) complex $[PhB(CH_2P(^{i}Pr)_2)_3]FeN$ (where Ph = phenyl and $^{i}$Pr = isopropyl), which shows an Fe≡N triple bond stretch at 1034 $cm^{-1}$ (*16*). The isomer shift of **5** is almost perfectly reproduced in the calculation, which supports our assignment of the oxidation state of **5** and also lends confidence to the calculated electronic structure detailed below. The sign of $\Delta E_Q$ is accurately calculated, and the magnitudes of the experimental and calculated values are consistent with a $d^2$ ion with a doubly occupied nonbonding $d_{xy}$ orbital and a strongly covalent iron-nitrogen multiple bond.

We were particularly interested in analyzing the nature of the Fe≡N bond in the context of the evolving understanding of metal-ligand multiple bonds (*40*, *41*). On the basis of our DFT electronic structure calculations (Fig. 4, right), the highest occupied molecular orbital (HOMO) is best described as a doubly occupied $d_{xy}$ orbital. The *xz* and *yz* orbitals are highly destabilized in energy because of considerable contributions (45%) from p orbitals of the terminal nitrido ligand and are best considered to be Fe–N π* antibonding orbitals. Occupancy of these orbitals reduces the Fe–N bond order, as in the case of **2**, which exhibits an Fe–N bond order of 2.5 (*28*). The calculated bond order [from Löwdin analysis (*42*)] in **5** is 2.82, justifying the first-order description of **5** as an Fe(VI)-nitrido complex with an Fe≡N triple bond. This view is fully consistent with the decrease in the Fe≡N bond distances from 1.61 Å in **2** (*28*) to 1.57 Å in **5**, which is comparable to the Mn≡N triple bond (1.54 Å) observed in $[(cy\text{-}ac)MnN]^{+}$ (*39*) and is 0.08 Å longer than the calculated Fe≡N triple bond distance in the tetrahedral Fe(IV)-nitrido complex $[PhB(CH_2P(^{i}Pr)_2)_3]FeN$ (*16*) as expected from the difference in coordination number. For comparison, a typical Fe=O double bond length in Fe(IV)-oxo species is 1.65 Å (*37*, *43*, *44*).

Compound **5** is stable for months in frozen solution at 77 K and is stable for some time in fluid solution at –40°C. A Mössbauer spectrum of a solution thawed to –40°C for 10 min showed only 29% diminished intensity of the signal due to **5**, which remained the major species. The decomposition product evident in the spectrum could be assigned as a high-spin Fe(III) species because of its Mössbauer parameters. Thus, **5** readily gains three electrons.

The discovery of **5** has implications for biological systems for which high-valent iron intermediates have often been proposed. Although Fe(IV)-oxo species have been observed in some cases [(*45*–*47*); see also (*48*)], concrete evidence of an Fe(V) intermediate is lacking. Postulation of Fe(V) intermediates is often met with skepticism because very few genuine Fe(V) coordination complexes are known. The results reported here should support further consideration of proposed high-valent iron intermediates.

**References and Notes**


1. L. H. Bowen, E. De Grave, R. E. Vandenberghe, in *Mössbauer Spectroscopy Applied to Magnetism and Materials Science*, G. J. Long, F. Grandjean, Eds. (Plenum, New York, 1993), vol. 1, pp. 115–159.
2. E. Murad, J. H. Johnston, in *Mössbauer Spectroscopy Applied to Inorganic Chemistry*, G. J. Long, Ed. (Plenum, New York, 1987), vol. 2, pp. 507–582.
3. G. Klingelhöfer *et al.*, *Science* **306**, 1740 (2004).
4. L. A. Haskin *et al.*, *Nature* **436**, 66 (2005).
5. H. Beinert, R. H. Holm, E. Münck, *Science* **277**, 653 (1997).
6. A. Volbeda *et al.*, *Nature* **373**, 580 (1995).
7. J. W. Peters, W. N. Lanzilotta, B. J. Lemon, L. C. Seefeldt, *Science* **282**, 1853 (1998).
8. S. Shima, E. J. Lyon, R. K. Thauer, B. Mienert, E. Bill, *J. Am. Chem. Soc.* **127**, 10430 (2005).
9. M. Merkx *et al.*, *Angew. Chem. Int. Ed.* **40**, 2782 (2001).
10. M. Costas *et al.*, *J. Am. Chem. Soc.* **123**, 12931 (2001).
11. I. G. Denisov, T. M. Makris, S. G. Sligar, I. Schlichting, *Chem. Rev.* **105**, 2253 (2005).
12. A. Chanda *et al.*, *J. Inorg. Biochem.* **100**, 606 (2006).
13. J. Kaizer *et al.*, *J. Am. Chem. Soc.* **126**, 472 (2004).
14. D. Kumar, H. Hirao, L. Que, S. Shaik, *J. Am. Chem. Soc.* **127**, 8026 (2005).
15. C. C. Cummins, R. R. Schrock, *Inorg. Chem.* **33**, 395 (1994).
16. T. A. Betley, J. C. Peters, *J. Am. Chem. Soc.* **126**, 6252 (2004).
17. M. P. Jensen *et al.*, *J. Am. Chem. Soc.* **127**, 10512 (2005).
18. L. D. Slep *et al.*, *J. Am. Chem. Soc.* **125**, 15554 (2003).
19. C. M. Thomas, N. P. Mankad, J. C. Peters, *J. Am. Chem. Soc.* **128**, 4956 (2006).
20. K. Meyer, E. Bill, B. Mienert, T. Weyhermüller, K. Wieghardt, *J. Am. Chem. Soc.* **121**, 4859 (1999).
21. C. A. Grapperhaus, B. Mienert, E. Bill, T. Weyhermüller, K. Wieghardt, *Inorg. Chem.* **39**, 5306 (2000).
22. L. Delaude, P. Laszlo, *J. Org. Chem.* **61**, 6360 (1996).
23. V. K. Sharma, *Water Sci. Technol.* **49**, 69 (2004).
24. J. Q. Jiang, B. Lloyd, *Water Res.* **36**, 1397 (2002).
25. S. Licht, B. H. Wang, S. Ghosh, *Science* **285**, 1039 (1999).
26. F. Kazama, *FEMS Microbiol. Lett.* **118**, 345 (1994).
27. K. Nakamoto, *Coord. Chem. Rev.* **226**, 153 (2002).
28. N. Aliaga-Alcalde *et al.*, *Angew. Chem. Int. Ed.* **44**, 2908 (2005).
29. J. F. Berry, E. Bill, E. Bothe, T. Weyhermüller, K. Wieghardt, *J. Am. Chem. Soc.* **127**, 11550 (2005).
30. J. F. Berry *et al.*, *Inorg. Chem.* **45**, 2027 (2006).
31. In samples of **3**, a portion of the molecules undergo spin crossover to the low-spin state as the temperature is lowered, but this proportion never accounts for more than 50% of the sample.
32. U. Russo, G. J. Long, in *Mössbauer Spectroscopy Applied to Inorganic Chemistry*, G. J. Long, F. Grandjean, Eds. (Plenum, New York, 1989), vol. 3, pp. 289–329.
33. N. N. Greenwood, T. C. Gibb, *Mössbauer Spectroscopy* (Chapman and Hall, London, 1971).
34. The iron-nitride stretch is not intense enough to be observed in the spectrum, as has been previously reported for **2** (*28*).
35. See supporting material on *Science* Online.
36. M. R. Bukowski *et al.*, *Science* **310**, 1000 (2005); published online 27 October 2005 (10.1126/science.1119092).

37. J. U. Rohde *et al.*, *J. Am. Chem. Soc.* **126**, 16750 (2004).
38. Because of the limited range in photoelectron wave vector space (or k-space), the resolution of the data is ~0.16 Å, and therefore the Fe–N(eq) and Fe–O(ax) bond distances could not be separated. We note, however, that the observed Fe–N/O bond distance of 2.03 ± 0.02 Å agrees well with what is expected for the average of four Fe–N bond distances of 2.06 Å and one Fe–O bond distance of 1.92 Å. The Fe≡N distance is considerably shorter than these distances and is therefore well resolved.
39. C. A. Grapperhaus, E. Bill, T. Weyhermüller, F. Neese, K. Wieghardt, *Inorg. Chem.* **40**, 4191 (2001).
40. V. M. Miskowski, H. B. Gray, M. D. Hopkins, in *Advances in Transition Metal Coordination Chemistry*, C.-M. Che, V. W. W. Yam, Eds. (JAI Press, Greenwich, CT, 1996), vol. 1, pp. 159–186.
41. C. C. Cummins, *Angew. Chem. Int. Ed.* **45**, 862 (2006).
42. P.-O. Löwdin, *Adv. Quantum Chem.* **5**, 185 (1970).
43. J. U. Rohde *et al.*, *Science* **299**, 1037 (2003).
44. E. J. Klinker *et al.*, *Angew. Chem. Int. Ed.* **44**, 3690 (2005).
45. J. C. Price, E. W. Barr, B. Tirupati, J. M. Bollinger, C. Krebs, *Biochemistry* **42**, 7497 (2003).
46. I. Schlichting *et al.*, *Science* **287**, 1615 (2000).
47. M. Newcomb *et al.*, *J. Am. Chem. Soc.* **128**, 4580 (2006).
48. R. Davydov *et al.*, *J. Am. Chem. Soc.* **123**, 1403 (2001).


49. Supported by the Fonds der Chemischen Industrie and by an Alexander von Humboldt Foundation postdoctoral fellowship (J.F.B.). The Stanford Synchrotron Radiation Laboratory is funded by the U.S. Department of Energy (DOE), Office of Basic Energy Sciences. The Structural Molecular Biology program is supported by NIH, National Center for Research Resources, Biomedical Technology Program, and DOE, Office of Biological and Environmental Research.


**Supporting Online Material**

www.sciencemag.org/cgi/content/full/1128506/DC1
Materials and Methods
Tables S1 and S2
Fig. S1
References


10 April 2006; accepted 19 May 2006
Published online 1 June 2006;
10.1126/science.1128506
Include this information when citing this paper.


# Molecular Recognition in the Selective Oxygenation of Saturated C-H Bonds by a Dimanganese Catalyst


**Siddhartha Das, Christopher D. Incarvito, Robert H. Crabtree,* Gary W. Brudvig***



**Although enzymes often incorporate molecular recognition elements to orient substrates selectively, such strategies are rarely achieved by synthetic catalysts. We combined molecular recognition through hydrogen bonding with C-H activation to obtain high-turnover catalytic regioselective functionalization of $sp^3$ C-H bonds remote from the –COOH recognition group. The catalyst contains a $Mn(\mu\text{-}O)_2Mn$ reactive center and a ligand based on Kemp's triacid that directs a –COOH group to anchor the carboxylic acid group of the substrate and thus modify the usual selectivity for oxidation. Control experiments supported the role of hydrogen bonding in orienting the substrate to achieve high selectivity.**


Selective functionalization of C-H bonds at sites in molecules remote from more reactive substituents is a major challenge in synthetic chemistry (*1*). In biological systems, monooxygenases and fatty acid desaturases have long been known to combine molecular recognition with reactive catalytic metal centers and thereby oxidize hydrocarbons with very high regio- and stereoselectivity (*2*–*9*). Multiple amino acid residues interact with the substrate via noncovalent interactions, such as hydrogen bonding and π-stacking, to position a specific reaction site in a precisely favorable orientation relative to a (typically iron-based) catalytic center. The ubiquitous hydroxylating cytochrome P450–based enzymes, for example, having an iron-porphyrin in the active site, constitute the most widely distributed class of regioselective monooxygenases (*2*, *3*). Noncovalent recognition is important to ensure adequate lability for catalytic turnover. This strategy is hard to implement in synthetic systems that, unlike proteins, must rely on much smaller, simpler scaffolding structures to influence substrate orientation. In one case, Breslow and co-workers grafted porphyrin-based manganese catalysts to cyclodextrin groups, which anchored substrates in a favorable orientation through hydrophobic interactions (*10*–*13*). Regioselective oxidation has also been achieved with other catalysts, but in this case substrates were aligned through covalent bonds, which prevented turnover (*14*–*16*).

Here we report a nonporphyrin di-μ-oxo dimanganese compound that catalyzes the highly regioselective oxygenation of saturated C-H bonds in ibuprofen and (4-methylcyclohexyl) acetic acid (cis + trans) with >100 turnovers. The selectivity arises from noncovalent molecular recognition via reversible H-bonding between a carboxylic acid group of the catalyst and a carboxylic acid group of the substrate.

The complex $[H_2O(L)Mn(\mu\text{-}O)_2Mn(L)OH_2](NO_3)_3$ (**1a**, L is 2,2′:6′,2″-terpyridine) has earlier been reported as a very active catalyst for oxidation chemistry with Oxone (peroxomonosulfate) as primary oxidant (*17*, *18*). Good evidence was obtained in this system to exclude the intermediacy of freely diffusing radicals, which would otherwise degrade the selectivity (*19*). Our present study uses the same di-μ-oxo dimanganese core, but with a different ligand, **2** (Fig. 1). Our ligand design addressed several key criteria: the need to stabilize high-valent manganese, an oxidation-resistant framework, and a –COOH group properly directed for molecular recognition. In ligand **2**, the terpy group was chosen to stabilize high-valent manganese (*20*, *21*). The phenylene linker provides a spacer between the docking element and the remote site of functionalization. The Kemp's triacid fragment provides a U-turn motif, a –COOH group suitably oriented for the molecular recognition function. The structure of the ligand was confirmed by x-ray crystallography (Fig. 1) (*22*). A suitable catalyst precursor, [(**2**)$MnCl_2$], was prepared by refluxing a solution of ligand **2** in acetonitrile and a saturated aqueous solution of

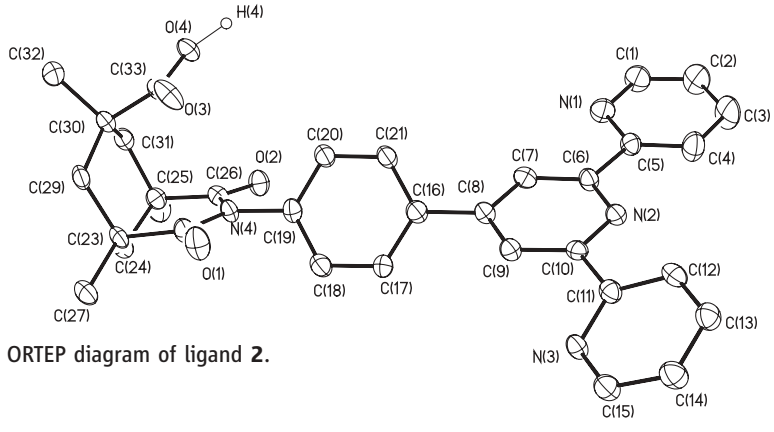



**Fig. 1.** ORTEP diagram of ligand **2**.


Department of Chemistry, Yale University, 225 Prospect Street, Post Office Box 208107, New Haven, CT 06520–8107, USA.

*To whom correspondence should be addressed. E-mail: robert.crabtree@yale.edu (R.H.C.); gary.brudvig@yale.edu (G.W.B.)

excess $MnCl_2$. Subsequent oxidation of [(**2**)$MnCl_2$] with 0.80 eq of Oxone in $CH_3CN$-$H_2O$ mixture (1:1) gave [$H_2O$(**2**)Mn(μ-$O)_2$Mn(**2**)$OH_2$]$(NO_3)_4$ (**1b**) (*22*, *23*).

Molecular modeling allowed us to predict which C-H bond in the substrate would be expected to come closest to the active site. The geometry of the proposed H-bonded catalyst-substrate complex was obtained by importing the crystal structure parameters of the ligand **2** (Fig. 1) and the di-μ-oxo dimanganese core into the model (*17*, *18*, *23*), docking ibuprofen, and then minimizing the energy (Fig. 2) (MM2, CAChe 5). We chose ibuprofen [2-(4-isobutyl-phenyl)-propionic acid] (**3**) as an appropriately rigid substrate with two sites of attack (indicated by bold arrows in Fig. 3). Oxidation occurs at the remote benzylic carbon to give **4** and at the vicinal position to give **5** (Fig. 3). If the catalysis operates via the catalyst-substrate complex predicted by the model (Fig. 2), then **4** should be the major product, with any initial –CH(OH) intermediate being rapidly oxidized further to the ketone (Fig. 3). With another rigid substrate, (4-methylcyclohexyl) acetic acid (**6c**: cis isomer; **6t**: trans isomer; **6**: **6c** + **6t**) docked via hydrogen bonding to **1b**, the model (fig. S17) suggests that the remote tertiary C-H bond at C6 (indicated by the solid arrows in Fig. 3) should undergo oxidation preferentially to give the corresponding alcohol **7** (*22*).

Catalytic runs were carried out with 1 eq of substrate, 0.001 eq of **1b**, and 5 eq of Tetrabutylammonium Oxone in acetonitrile. At room temperature, the reaction was quenched after 2 hours by addition of excess $NaHSO_3$; products and unreacted substrate were extracted into ether. To confirm the role of docking, we performed control experiments using [$H_2O$(L′)Mn(μ-$O)_2$Mn(L′)$OH_2$]$(ClO_4)_3$ (**1c**, L′ is 4′-phenyl-2,2′:6′,2″-terpyridine) (*23*), a catalyst lacking the key –COOH group; under identical conditions, selectivity was lost. We also confirmed that no Oxone-induced reactivity occurs in the absence of the manganese catalyst (*22*).

When ibuprofen was treated with catalyst **1c**, which lacks the key –COOH group, the ratio of remote to proximal oxidation products (**4**:**5**) was roughly 3:1. With catalyst **1b**, however, the recognition functionality raised the selectivity for **4** more than 10-fold (Table 1). In time-dependent studies, the ratios of the products remained constant, confirming that the regioselectivity is not due to degradation of the alternative product **5** (*22*, *24*, *25*).

We ascribe this high regioselectivity to the modeled docking H-bonding between the –COOH groups of ibuprofen and catalyst **1b**. To test this hypothesis further, we added acetic acid to the reaction medium, expecting that it would largely displace the ibuprofen from the recognition site. If so, the catalytic selectivity would then resemble that observed in the absence of recognition. Indeed, the **4**:**5** ratio found in this case fell precisely to the level observed with the control catalyst **1c** (Table 1).

Oxidation of the alkyl carboxylic acid substrate **6**, using recognition catalyst **1b**, led not only to regioselective oxygenation at the remote tertiary C-H bond, but also to diastereoselection of a single isomer of **7**: the trans isomer **7t** (Table 2). Our control experiment with catalyst **1c** yielded several other oxidation products in addition to **7t** (*22*). This diastereoselectivity can be rationalized on the basis of a model (Fig. 4) in which the stereochemical outcome is determined in the "rebound" step. In this model, the carbon radical putatively formed by H-atom abstraction from the substrate adopts a chairlike conformation (*26*), which leads to the observed product isomer **7t** on transfer of an OH group from the Mn-OH intermediate to the carbon radical center. Transfer of OH to the opposite side of the ring, to yield the other product isomer, is not possible via any plausible intermediate conformations that we could identify from modeling work.

In a control experiment that rules out undesirable autoxidative mechanisms involving atmospheric molecular oxygen, the same results were obtained under a nitrogen

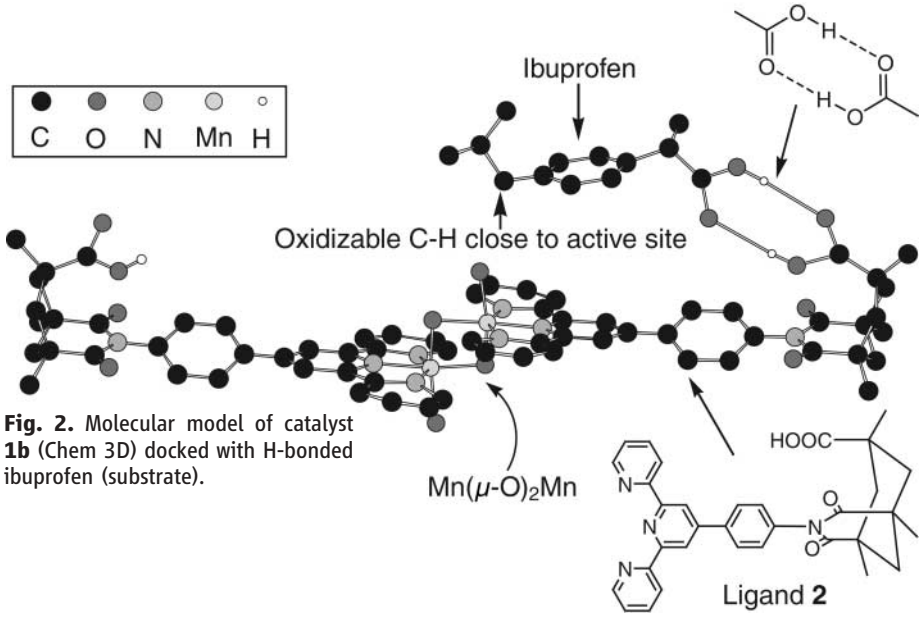



**Fig. 2.** Molecular model of catalyst **1b** (Chem 3D) docked with H-bonded ibuprofen (substrate).

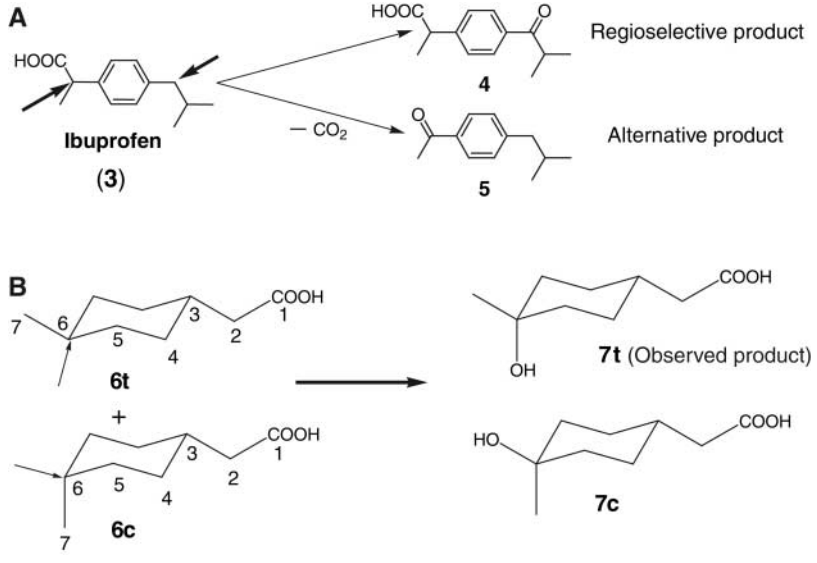



**Fig. 3.** (**A**) Oxidation products of ibuprofen with **1b** as catalyst. Bold arrows indicate alternate sites of attack. (**B**) Oxidation products of **6** (**6c** + **6t**) with **1b** as catalyst. Solid arrows indicate possible sites of attack according to molecular modeling.

atmosphere as were obtained in air. Any rebound must, therefore, be fast because the proposed carbon radical is not trapped by ambient molecular oxygen. Oxidation of *cis*-stilbene yielded *cis*-stilbene epoxide, not the *trans* epoxide, which is usually the major product from a freely diffusing radical mechanism. This result suggests that a freely diffusing radical mechanism is unlikely (*19*, *27*). Further mechanistic studies are in progress.

Initially, we obtained low catalytic turnovers (~50) for ibuprofen oxidation using 1% catalyst (relative to substrate), probably because catalyst degradation halted the reaction. To remedy this problem, we lowered the catalyst concentration, thereby reducing the probability of bimolecular catalyst self-oxidation. With a 0.1% catalyst to substrate ratio, a much higher total turnover number of 580 was attained (catalyst **1b**, 0°C) with no loss of regioselectivity (98.5%) (Table 1). For oxidation of **6**, we used 0.1% catalyst in all experiments. Further improvement of the catalytic turnover has been attained by replacing the $CH_3CN$ solvent with more oxidation-resistant $CD_3CN$, thereby reducing the competitive solvent oxidation. Use of $CD_3CN$ raised total turnover numbers even further without adversely affecting the regioselectivity (and diastereoselectivity in the case of **6**) (Tables 1 and 2).

**Table 1.** Product distribution from ibuprofen oxidation [by $^1$H–nuclear magnetic resonance ($^1$H-NMR) spectroscopy].

| Temperature | Catalyst | Conversion | 4 (Favored by recognition) | 5 (Disfavored by recognition) | Total turnovers* |
|---|---|---|---|---|---|
| 20°C | **1b** | 50% | 97.5% | 2.5% | 50 |
| | **1c** | 53% | 77% | 23% | 53 |
| 0°C | **1b** | 53% | 98.5% | 1.5% | 53 |
| | **1c** | 54% | 78% | 22% | 54 |
| –20°C | **1b** | 53% | 98.5% | 1.5% | 53 |
| | **1c** | 54% | 77% | 23% | 54 |
| 20°C | **1b**† | 56% | 75% | 25% | 56 |
| | **1c**† | 58% | 77% | 23% | 58 |
| 0°C | **1b**§ | 58% | 98.5% | 1.5% | 580 |
| 20°C | **1b**‡§ | 71% | 96.5% | 3.5% | 710 |

*Total turnovers = mol products per mol catalyst; subtrate:catalyst:oxidant = 100:1:500. †Solutions contained excess acetic acid (400% with respect to substrate). ‡With $CD_3CN$ as solvent instead of $CH_3CN$. §Subtrate:catalyst:oxidant = 100:0.1:500.

**Table 2.** Product distribution from **6** (**6c** + **6t**) (by $^1$H-NMR spectroscopy).

| Temperature | Catalyst (0.1%) | Conversions | 7t (favored) | 7c (disfavored) | Other products | Total turnovers* |
|---|---|---|---|---|---|---|
| 20°C | **1b** | 13% | >99% | <1% | <1% | 130 |
| 20°C | **1c** | ~19% | ~30% | ~30% | ~40% | 190 |
| 20°C | **1b**† | 18% | >99% | <1% | <1% | 180 |

*Total turnovers = mol products per mol catalyst; subtrate:catalyst:oxidant = 100:0.1:500. †With $CD_3CN$ as solvent instead of $CH_3CN$.

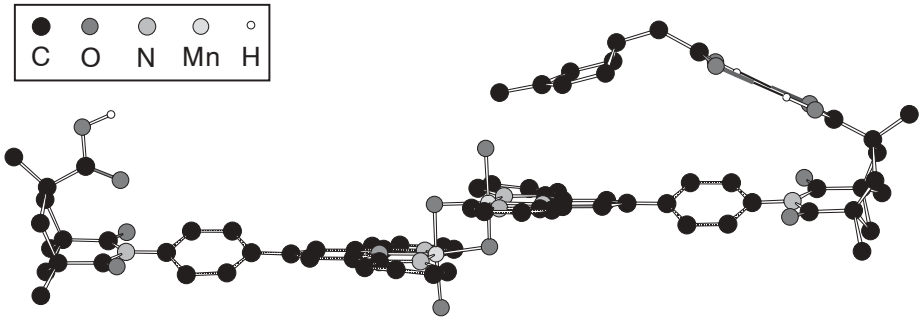



**Fig. 4.** Molecular model of the intermediate, resulting from H-atom abstraction from C6 of **6** in a distorted chair conformation, docked to **1b** (Chem 3D).

The general strategy may have wide application in chemical catalysis.

### References and Notes


1. B. Meunier, Ed., *Biomimetic Oxidations Catalyzed by Transition Metal Complexes* (Imperial College Press, London, 2000).
2. P. R. Ortiz de Montellano, Ed., *Cytochrome P450: Structure, Mechanism, and Biochemistry* (Plenum, New York, ed. 2, 1995).
3. J. H. Dawson, M. Sono, *Chem. Rev.* **87**, 1255 (1987).
4. B. Behrouzian, P. H. Buist, *Prostaglandins Leukot. Essent. Fatty Acids* **68**, 107 (2003).
5. J. B. van Beilen, J. Kingma, B. Witchold, *Enzyme Microbiol. Technol.* **106**, 7928 (1994).
6. B. Behrouzian, C. K. Savile, B. Dawson, P. H. Buist, J. Shanklin, *J. Am. Chem. Soc.* **124**, 3277 (2002).
7. M. Moche, J. Shanklin, A. Ghoshal, Y. Lindqvist, *J. Biol. Chem.* **278**, 25072 (2002).
8. J. A. Broadwater, E. Whittle, J. Shanklin, *J. Biol. Chem.* **277**, 15613 (2002).
9. B. G. Fox, K. S. Lyle, C. E. Rogge, *Acc. Chem. Res.* **37**, 421 (2004).
10. R. Breslow, X. Zhang, Y. Huang, *J. Am. Chem. Soc.* **119**, 4535 (1997).
11. R. Breslow, Y. Huang, X. Zhang, J. Yang, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 11156 (1997).
12. J. Yang, R. Breslow, *Angew. Chem. Int. Ed. Engl.* **39**, 2692 (2000).
13. J. Yang, B. Gabriele, S. Belvedere, Y. Huang, R. Breslow, *J. Org. Chem.* **67**, 5057 (2002).
14. R. F. Moreira, P. M. Wehn, D. Sames, *Angew. Chem. Int. Ed. Engl.* **39**, 1618 (2000).
15. See (*28*) for an example of chelate-directed C-H bond activation.
16. M. D. Kaufman, P. A. Grieco, D. W. Bougie, *J. Am. Chem. Soc.* **115**, 11648 (1993).
17. J. Limburg *et al.*, *Science* **283**, 1524 (1999).
18. J. Limburg *et al.*, *J. Am. Chem. Soc.* **123**, 423 (2001).
19. J. Wessel, R. H. Crabtree, *J. Mol. Catal. A* **113**, 13 (1996).
20. Terpy moiety has been used previously for molecular recognition studies (*29*).
21. M. S. Goodman, V. Jubian, B. Linton, A. D. Hamilton, *J. Am. Chem. Soc.* **117**, 11610 (1995).
22. See supporting online material.
23. H. C. Chen *et al.*, *Inorg. Chem.* **44**, 7661 (2005).
24. G. Caviglioli *et al.*, *J. Pharm. Biomed. Anal.* **30**, 499 (2002).
25. M. Komuro, Y. Nagatsu, T. Higuchi, M. Hirobe, *Tetrahedron Lett.* **33**, 4949 (1992).
26. The chairlike conformation (as in Fig. 4) of 1-methyl-1-cyclohexyl cation has been reported to be more stable than its ideal chair conformation (*30*).
27. A. J. Castellino, T. C. Bruice, *J. Am. Chem. Soc.* **110**, 158 (1988).
28. B. D. Dangel, J. A. Johnson, D. Sames, *J. Am. Chem. Soc.* **123**, 8149 (2001).
29. M. S. Goodman, A. D. Hamilton, J. Weiss, *J. Am. Chem. Soc.* **117**, 8447 (1995).
30. A. Rauk *et al.*, *J. Am. Chem. Soc.* **118**, 3761 (1996).


31. Supported by the NIH grant GM32715. We thank L. Que Jr. for suggestions; A. D. Hamilton for use of his high-performance liquid chromatograph; and J. W. Faller for use of his CAChe 5 program. Full data for the crystal structure of ligand **2** are available free of charge from Cambridge Crystallographic Data Center under reference number CCDC 606493


### Supporting Online Material

www.sciencemag.org/cgi/content/full/312/5782/1941/DC1
Materials and Methods
SOM Text
Figs. S1 to S17
References

27 March 2006; accepted 9 May 2006
10.1126/science.1127899

# The Competitive Cost of Antibiotic Resistance in *Mycobacterium tuberculosis*

Sebastien Gagneux,[1,4*†] Clara Davis Long,[2*] Peter M. Small,[4,5] Tran Van,[1] Gary K. Schoolnik,[1,3] Brendan J. M. Bohannan[2]

Mathematical models predict that the future of the multidrug-resistant tuberculosis epidemic will depend on the fitness cost of drug resistance. We show that in laboratory-derived mutants of *Mycobacterium tuberculosis*, rifampin resistance is universally associated with a competitive fitness cost and that this cost is determined by the specific resistance mutation and strain genetic background. In contrast, we demonstrate that prolonged patient treatment can result in multidrug-resistant strains with no fitness defect and that strains with low- or no-cost resistance mutations are also the most frequent among clinical isolates.

Antimicrobial resistance has become a worldwide problem in many pathogens, incurring both economic costs and loss of human lives (*1*, *2*). Antibiotic resistance is also often associated with a reduced competitive ability against antibiotic-sensitive strains, in the absence of the antibiotic (*3*). In a variety of model systems, it has been shown that this fitness "cost" depends on the specific drug resistance–conferring mutation and the strain genetic background and that it can be ameliorated by compensatory mutations (*4*–*9*). However, there are few studies linking the predictions of these model systems to epidemiological data from human populations (*10*–*12*). Here, we connect the in vitro cost of resistance to the prevalence of clinically important mutants of *Mycobacterium tuberculosis*, the causative agent of human tuberculosis.

The emergence of multidrug-resistant strains of *M. tuberculosis* is threatening global disease control efforts (*13*). Multidrug-resistant tuberculosis (MDRTB) is defined as tuberculosis caused by organisms resistant to isoniazid and rifampin, the two most potent first-line antituberculosis drugs. MDRTB represents a major public health burden, requiring prolonged treatment with more costly and less effective agents (*13*, *14*). To date, most investigations of factors determining the spread of MDRTB have focused on the quality of tuberculosis-control programs and patient adherence to treatment. However, recent studies suggest that bacterial factors, such as the relative fitness of drug-resistant *M. tuberculosis* strains, are also important (*15*–*18*). *M. tuberculosis* develops drug resistance by the sequential acquisition of chromosomal mutations (*19*), but there are only limited data on the effect of different drug resistance–conferring mutations on the relative fitness of *M. tuberculosis* (*12*, *20*). Furthermore, there has been no attempt to investigate the effects of different genetic backgrounds on the competitive fitness of drug-resistant pathogens, including *M. tuberculosis.*

The relative competitive fitness of bacteria can be quantified experimentally with the use of competition assays originally developed for *Escherichia coli* (*21*). In these experiments, the drug-susceptible and the drug-resistant organisms compete for limited resources in a common environment. In *M. tuberculosis*, as in other bacteria, resistance to rifampin is mediated through missense mutations in *rpoB*, the gene encoding the β subunit of the RNA polymerase (*19*). To determine the impact of different *rpoB* mutations on the relative competitive fitness of rifampin-resistant *M. tuberculosis*, we selected a panel of spontaneous rifampin-resistant mutants by growing *M. tuberculosis* CDC1551 on media containing rifampin. CDC1551 is a clinical strain and part of a lineage of *M. tuberculosis* associated with Europe and the Americas (*22*–*24*), and it is susceptible to currently used antituberculosis drugs. We sequenced known rifampin-resistance regions in the *rpoB* gene of all of the 52 colonies recovered and found that all of them had a missense substitution in *rpoB* (table S1). We chose nine mutants with different *rpoB* mutations and had them compete in vitro against their rifampin-susceptible ancestor, with the use of an assay that included lag, exponential, and early stationary phases of growth (*21*). All of the rifampin-resistant mutants had a statistically significant fitness cost compared with the susceptible ancestor, which is defined to have a competitive fitness of 1.0 (Fig. 1 and table S2). Among the different *rpoB* mutants, there were significant differences in competitive fitness, with values ranging from 0.58 to 0.91. We found a significant effect of mutation on fitness [analysis of variance (ANOVA), $F_{8,38} = 12.1563$, $P < 0.0001$]. The $Ser^{531} \rightarrow Leu^{531}$ (S531L) mutant exhibited the lowest cost, whereas the $Arg^{529} \rightarrow Gln^{529}$ (R529Q) mutant had the greatest fitness cost of all mutants (Fig. 1 and table S2). These results confirm that a wide range of *rpoB* mutations can have markedly different effects on the competitive fitness of *M. tuberculosis* in vitro, consistent with earlier work that examined the competitive cost of a small number of *rpoB* mutations (*12*, *20*).

Next, we investigated the influence of the bacterial genetic background on the cost of rifampin resistance–conferring mutations. In the laboratory, we selected a second panel of spontaneous rifampin-resistant mutants in T85, another pan-susceptible isolate that is part of a distinct lineage of *M. tuberculosis* strongly associated with East Asia, commonly referred to as the Beijing strain (*24*, *25*). All 63 rifampin-resistant colonies selected in T85 had a nonsynonymous substitution in *rpoB* (table S3). We measured the relative competitive fitness of four different mutants with *rpoB* mutations whose fitness costs had been measured in CDC1551. Similar to the results in the CDC1551 back-

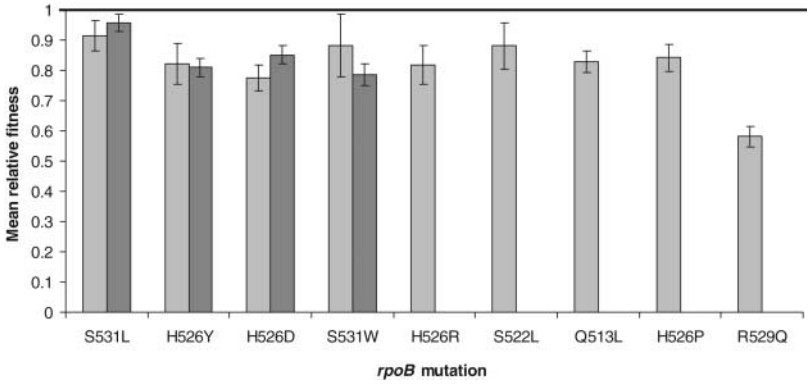



**Fig. 1.** Relative competitive fitness of laboratory-derived rifampin-resistant mutants of *M. tuberculosis*. All mutants had a statistically significant fitness cost (error bars indicate 95% confidence intervals). This cost was less in *rpoB* S531L mutants than in other *rpoB* mutants, irrespective of the strain background. Light gray bars, CDC1551 mutants; dark gray bars, T85 mutants. Y, Tyr; W, Trp; P, Pro.

[1]Division of Infectious Diseases and Geographic Medicine, [2]Department of Biological Sciences, [3]Department of Microbiology and Immunology, Stanford University, Stanford, CA 94305, USA. [4]Institute for Systems Biology, Seattle, WA 98103, USA. [5]Bill and Melinda Gates Foundation, Seattle, WA 98102, USA.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: sgagneux@systemsbiology.org

ground, all T85 rifampin-resistant mutants had a significant fitness cost (Fig. 1 and table S2). We tested for an effect of genetic background by comparing the fitness costs of the four mutations found in each of the two backgrounds. We found an overall significant effect of genetic background when nested in mutation (nested ANOVA, $F_{7,34} = 6.8962$, $P < 0.0001$), and, as observed in the CDC1551 background, the *rpoB* S531L mutant had higher competitive ability than other mutations in the T85 background (Fig. 1 and table S2). The fitness of the S531L mutation did not significantly differ in the two genetic backgrounds (i.e., the confidence intervals overlap; Fig. 1 and table S2). However, genetic background did significantly alter the fitness of the $His^{526} \rightarrow Asp^{526}$ (H526D) mutation (i.e., the confidence intervals do not overlap; Fig. 1 and table S2). Taken together, our data confirm that the in vitro relative fitness of rifampin-resistant *M. tuberculosis* depends on both the resistance-conferring mutation and the strain genetic background.

The relevance of empirically measured costs of laboratory-derived drug resistance to resistance that emerges in tuberculosis patients has not been established. Therefore, we measured the relative competitive fitness of rifampin-resistance in paired isolates (one that was drug resistant and one that was drug susceptible) from 10 patients who acquired rifampin resistance during antibiotic treatment (Fig. 2 and table S4). We used a standard DNA fingerprinting method to confirm that the paired isolates belonged to the same bacterial clone (*26*). In all 10 patients, both isolates had identical DNA fingerprinting patterns (table S5), which supports the assertion that the rifampin-resistant isolates recovered later were indeed derived from the corresponding rifampin-sensitive isolates. However, DNA fingerprinting methods are limited and may not reveal small genetic differences which could have been present between isolates. All of the clones were distinct but belonged to one of the two strain lineages represented by CDC1551 and T85 (tables S4 and S5). Five of the resistant isolates harbored the *rpoB* S531L mutation and five had other *rpoB* mutations (Fig. 2, table S4).

All of the clinical isolates with mutations other than S531L had a relative fitness of less than 1.0, although the measured relative fitness was not always the same as that observed in the mutants selected in vitro (Fig. 2 and table S4). In contrast, four of the five clinical strains with the *rpoB* S531L mutation had a relative fitness of greater than or equal to 1.0. Notably, these four strains were the only rifampin-resistant mutants that had no fitness cost compared with their rifampin-susceptible ancestor.

Two alternative explanations could account for the high fitness we observed in clinical strains harboring the *rpoB* S531L mutation. First, this mutation may indeed be a no-cost (versus just a low-cost) mutation in these particular strains. Second, given that sufficient time had passed between the isolation of the susceptible and resistant members of each pair (table S4), compensatory mutations could have arisen and ameliorated the low initial fitness cost associated with this mutation, similar to what has been observed in other species of bacteria (*4*, *5*, *8*). In support of this possibility, we found that the clinically derived *rpoB* S531L mutants had a mean competitive fitness of 1.04 (95% confidence interval: 1.00 to 1.08). This was statistically significantly higher than the laboratory-derived *rpoB* S531L mutants, which had a mean competitive fitness of 0.93 (95% confidence interval: 0.90 to 0.96). Given that all clinical ancestors of the *rpoB* S531L mutants were already resistant to isoniazid or isoniazid and streptomycin (table S4), this increase in fitness could be due to amelioration of fitness costs associated with the other resistance mutations rather than specifically the rifampin-resistance mutation. In addition, general increases in fitness could have occurred during long-term infection, rather than specific amelioration of the cost associated with a resistance mutation. We cannot differentiate between these alternative hypotheses with our current data. Nevertheless, our results clearly demonstrate that clinical strains of multidrug-resistant *M. tuberculosis* can have a significantly higher fitness than their progenitors, either because of no-cost resistance mutations, compensatory evolution, or both.

We hypothesized that those drug-resistant strains with the least costly and potentially most easily compensated resistance mutation will be more likely to spread and become prevalent in human populations. This scenario is particularly likely given the recent observation that many patients harbor subpopulations of the same mycobacterial clone with different drug resistance–conferring mutations (*27*). Many studies have reported the frequency of different drug-resistance alleles in a variety of clinical settings (*28*), providing an opportunity to test our hypothesis. When we compared the relative fitness of the different *rpoB* mutants to their clinical frequency, an association became evident (Figs. 1 and 2 and tables S2 and S4). The *rpoB* S531L mutation, which exhibited the lowest fitness cost in laboratory-derived mutants and no fitness cost in clinical strains, is the most prevalent rifampin resistance–conferring mutation, overall accounting for 54% of rifampin-resistant isolates, even in parts of the world associated with phylogenetically distinct strain lineages (*24*, *28*). In contrast, the *rpoB* R529Q mutant, which carried the highest fitness cost of all mutants, has never been observed in clinical settings (*28*). Taken together, these findings are consistent with the idea that drug-resistant strains harboring low- or no-cost mutations such as *rpoB* S531L are selected in patients during treatment and that such strains are more likely to spread in human populations.

Several mathematical models have been developed to predict the future of MDRTB epidemics. Some have assumed a universal fitness cost to drug resistance and concluded that MDRTB will remain a localized problem (*15*). More recent mathematical models have allowed for variable fitness among strains and have come to very different conclusions depending on the specific fitness assumed (*17*). Our work supports the more complex models and suggests that the heterogeneity in fitness is a function of the drug-resistance mutation, the strain genetic background, and potentially compensatory evolution as well. These more complex models suggest that resistance will not spread below a relative fitness of 0.7. Interestingly, all the *rpoB* mutants tested had a relative fitness >0.7, except for the one that has not been observed in clinical settings (Figs. 1 and 2). None of these models has allowed for changing fitness through time, either as a result of compensatory evolution or other increases in fitness

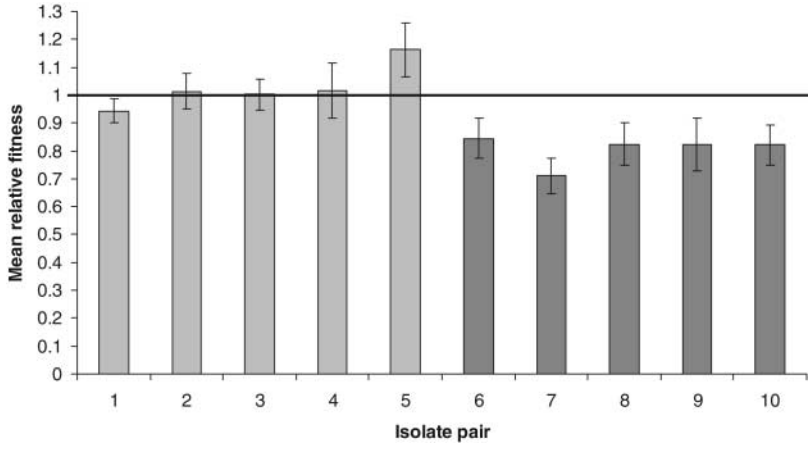



**Fig. 2.** Relative competitive fitness of clinically derived rifampin-resistant mutants of *M. tuberculosis*. Four of the five mutants with the *rpoB* S531L mutation (light gray bars) had no fitness cost compared with their rifampin-susceptible ancestors. All mutants with other *rpoB* mutations (dark gray bars) had significant fitness defects (error bars indicate 95% confidence intervals).

due to within-host evolution. Future mathematical treatments of the MDRTB epidemic should include such changes.

Our work suggests that in vitro competition assays can be predictive of biological differences important to *M. tuberculosis* ecology and evolution. Our finding that genetic background is a factor in the measured cost of resistance suggests that the cost of a given mutation may vary from location to location worldwide owing to the phylogeography of this pathogen (*24*). The possibility that fitness changes can occur during infection of a single patient should be taken into account when applying evolutionary ecology to infectious disease treatments or to the prediction of epidemics. The availability of such predictive methodology permits the incorporation of fitness considerations into the design and deployment of new drug regimens for controlling tuberculosis.

**References and Notes**


1. S. R. Palumbi, *Science* **293**, 1786 (2001).
2. T. M. File Jr., *Chest* **115**, 3S (1999).
3. D. I. Andersson, B. R. Levin, *Curr. Opin. Microbiol.* **2**, 489 (1999).
4. M. G. Reynolds, *Genetics* **156**, 1471 (2000).
5. S. Maisnier-Patin, D. I. Andersson, *Res. Microbiol.* **155**, 360 (2004).
6. F. M. Cohan, E. C. King, P. Zawadzki, *Evol. Int. J. Org. Evol.* **48**, 81 (1994).
7. B. Bjorkholm *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 14607 (2001).
8. S. J. Schrag, V. Perrot, *Nature* **381**, 120 (1996).
9. E. C. Böttger, B. Springer, M. Pletschette, P. Sander, *Nat. Med.* **4**, 1343 (1998).
10. T. A. Wichelhaus *et al.*, *Antimicrob. Agents Chemother.* **46**, 3381 (2002).
11. P. Sander *et al.*, *Antimicrob. Agents Chemother.* **46**, 1204 (2002).
12. O. J. Billington, T. D. McHugh, S. H. Gillespie, *Antimicrob. Agents Chemother.* **43**, 1866 (1999).
13. World Health Organization, *Anti-Tuberculosis Drug Resistance in the World—Third Global Report*, G. T. C. Program, Ed. (World Health Organization, Geneva, 2004).
14. J. S. Mukherjee *et al.*, *Lancet* **363**, 474 (2004).
15. C. Dye, M. A. Espinal, *Proc. R. Soc. London Ser. B* **268**, 45 (2001).
16. S. M. Blower, T. Chou, *Nat. Med.* **10**, 1111 (2004).
17. T. Cohen, M. Murray, *Nat. Med.* **10**, 1117 (2004).
18. C. Dye, B. G. Williams, M. A. Espinal, M. C. Raviglione, *Science* **295**, 2042 (2002).
19. S. Ramaswamy, J. M. Musser, *Tuber. Lung Dis.* **79**, 3 (1998).
20. D. H. Mariam, Y. Mengistu, S. E. Hoffner, D. I. Andersson, *Antimicrob. Agents Chemother.* **48**, 1289 (2004).
21. R. S. Lenski, M. R. Rose, S. C. Simpson, S. C. Tadler, *Am. Nat.* **138**, 1315 (1991).
22. L. Baker, T. Brown, M. C. Maiden, F. Drobniewski, *Emerg. Infect. Dis.* **10**, 1568 (2004).
23. R. D. Fleischmann *et al.*, *J. Bacteriol.* **184**, 5479 (2002).
24. S. Gagneux *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 2869 (2006).
25. A. G. Tsolaki *et al.*, *J. Clin. Microbiol.* **43**, 3185 (2005).
26. P. F. Barnes, M. D. Cave, *N. Engl. J. Med.* **349**, 1149 (2003).
27. F. A. Post *et al.*, *J. Infect. Dis.* **190**, 99 (2004).
28. D. M. O'Sullivan, T. D. McHugh, S. H. Gillespie, *J. Antimicrob. Chemother.* **56**, 674 (2005); published online 6 April 2005 (10.1093/jac/dki069).


29. We thank K. Kremer and A. Ponce de Leon for providing clinical strains. This research was supported by the NIH and the Wellcome Trust. S.G. was supported by the Swiss National Science Foundation and the Novartis Foundation.


**Supporting Online Material**

www.sciencemag.org/cgi/content/full/312/5782/1944/DC1
Materials and Methods
Tables S1 to S5
References

28 December 2005; accepted 30 May 2006
10.1126/science.1124410

# Lhx2 Maintains Stem Cell Character in Hair Follicles

**Horace Rhee, Lisa Polak, Elaine Fuchs***

**During embryogenesis, stem cells are set aside to fuel the postnatal hair cycle and repair the epidermis after injury. To define how hair follicle stem cells are specified and maintained in an undifferentiated state, we developed a strategy to isolate and transcriptionally profile embryonic hair progenitors in mice. We identified Lhx2 as a transcription factor positioned downstream of signals necessary to specify hair follicle stem cells, but upstream from signals required to drive activated stem cells to terminally differentiate. Using gain- and loss-of-function studies, we uncovered a role for Lhx2 in maintaining the growth and undifferentiated properties of hair follicle progenitors.**

Hair follicle morphogenesis involves a temporal series of reciprocal interactions between the ectoderm and its underlying mesenchyme (fig. S1) (*1*–*3*). In response to an inductive Wnt and an inhibitory Bmp signal (Noggin), small hair placodes bud from the epithelium, giving rise to larger hair germs (*4*–*7*). In the presence of the mitogen Shh, these hair germs develop further and grow downward to form a mature follicle that actively produces hair (*8*–*10*). Although the molecular details of bud formation are still poorly defined, the general features of this process are repeated at the start of each postnatal hair cycle when multipotent stem cells in the hair follicle bulge become activated to initiate a new round of hair growth. In addition, the early epithelial remodeling to form the hair germ shares many features with the development of other epithelial tissues and organs, including feathers, teeth, and mammary glands (*11*–*13*). Understanding how tissues form buds that then progress along different lineages is predicated on elucidating the molecular mechanisms that funnel these early signaling pathways into a transcriptional program that drives morphogenesis.

To examine the genetic changes that occur during epithelial bud formation, we developed a strategy to isolate embryonic hair progenitors. To this end, we generated mice doubly transgenic for a *Keratin 14–GFP* gene expressed in skin keratinocytes and the Wnt reporter gene *TOPGAL*, transcribed in hair placodes and germs where β-catenin/Lef1 complexes are active (*4*, *14*). In these early hair progenitors, E-cadherin is down-regulated and P-cadherin is up-regulated (Fig. 1A) (*7*). By embryonic day 17 (E17), we could use dispase to separate the epidermis, including hair placodes and germs, from the underlying dermis, which harbored more mature hair pegs and follicles (fig. S2). With the use of fluorescence-activated cell sorting (FACS) on the epidermal fraction, the early "PCAD+" hair progenitors (K14-GFP+, α6-integrin+, P-cadherin+) were then separated from the "PCAD–" interfollicular epidermis (K14-GFP+, α6-integrin+, P-cadherin–) on the basis of their differential surface P-cadherin expression (fig. S3). Characterization of these two cell populations confirmed that they had similarities in the expression of K5 and β4-integrin but distinct activities of *TOPGAL* and the expression of known hair-placode markers (Fig. 1, B to D, and fig. S4).

The gene expression profiles of purified PCAD+ hair progenitors and PCAD– interfollicular basal keratinocytes were further analyzed using oligonucleotide microarrays. Using fold differences of known hair-placode markers as a sensitivity gauge, a twofold cutoff was assigned as a genuine difference between the two populations. A total of 1394 probes (660 in PCAD+ and 734 in PCAD–) were preferentially expressed in one population relative to the other (table S1).

A short list of differentially expressed genes that are relevant to the present study is provided in table S2. As anticipated, a number of these genes have documented roles in either hair morphogenesis (PCAD+) or epidermal differentiation (PCAD–). The interfollicular epidermal population was typified by adhesive and cytoskeletal components, Notch signaling factors, c-Myc, Kruppel-like factors, and Bmp-responsive transcription factors (Grainyhead-like and Ovo1) previously implicated in epidermal differentiation (*15*–*18*). In contrast, the hair germ signature featured Wnts, Shh, Bmps, transforming growth

Howard Hughes Medical Institute, The Rockefeller University, 1230 York Avenue, New York, NY 10021.

*To whom correspondence should be addressed. E-mail: fuchslb@rockefeller.edu

factor–β's, and tyrosine kinase receptor signaling morphogens, as well as a number of different transcription factors. Although some of these transcription factors have not been previously implicated in the specification of skin progenitor fates, others have previously been associated with postnatal genetic hair disorders, including Cutl1, Gli1, Hoxc13, Sox9, Trps1, and Vdr (*2*, *3*).

Several of the uncharacterized transcription factors on this list were also found to be differentially expressed in the postnatal hair follicle bulge (*19*, *20*) (table S2), suggesting that the embryonic hair germ may exhibit functional properties similar to those of adult stem cells. Although the hair germ is committed to a follicular cell fate, it remains undifferentiated like bulge stem cells, yet capable of differentiating into all the lineages of the hair follicle, including the sebaceous gland (*21*, *22*).

To explore the possibility that the early hair germs may reflect hair follicle stem cells and regulate key steps in progenitor cell differentiation, we focused on transcription factors emanating from our screen that are known to govern developmental cell fate specification in other tissues and organs. Lim-homeodomain transcription factor Lhx2 was particularly interesting because *Lhx2* null mutant animals display defects in patterning and cell fate determination during brain development (*23*–*25*). In addition, they lack definitive erythropoeisis and conversely, hematopoetic progenitor cells can be maintained in vitro by forced expression of Lhx2 (*26*). *Lhx2* null animals die between E15.5 and E16.5, and a possible role for Lhx2 in skin has not been examined.

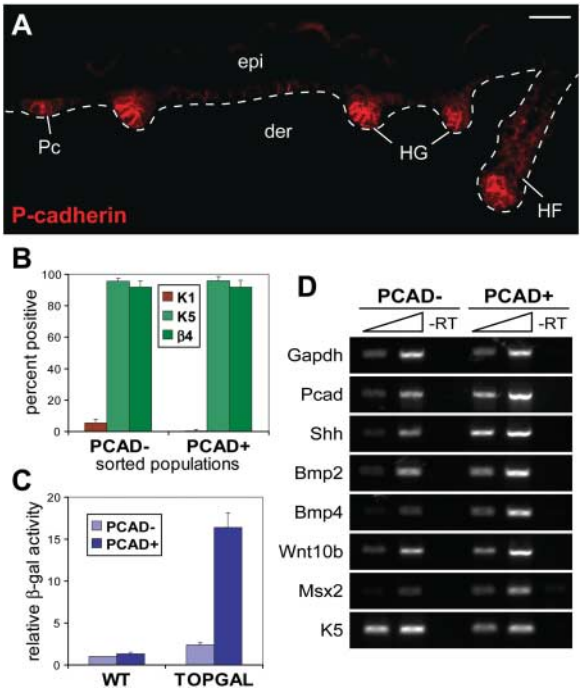



**Fig. 1.** Isolation of embryonic hair follicle progenitors. (**A**) P-cadherin is up-regulated at sites of hair follicle morphogenesis. This differential expression was used to isolate PCAD+ hair progenitors and PCAD– interfollicular basal cells by FACS (figs. S2 and S3). Scale bar, 40 μm. (**B**) Summary of cytospin analyses for epidermal markers K1, K5, and β4-integrin. (**C**) The differential activity of the Wnt reporter gene *TOPGAL* was assessed by chemiluminescence. β-gal, β-galactosidase; WT, wild type. (**D**) Semi-quantitative RT-PCR shows differential expression of known placode mRNAs. K5 and K14 are known to be down-regulated, but still expressed, in hair germs (*15*). Abbreviations are as follows: epi, epidermis; der, dermis; Pc, hair placode; HG, hair germ; HF, hair follicle; β4, β4-integrin.

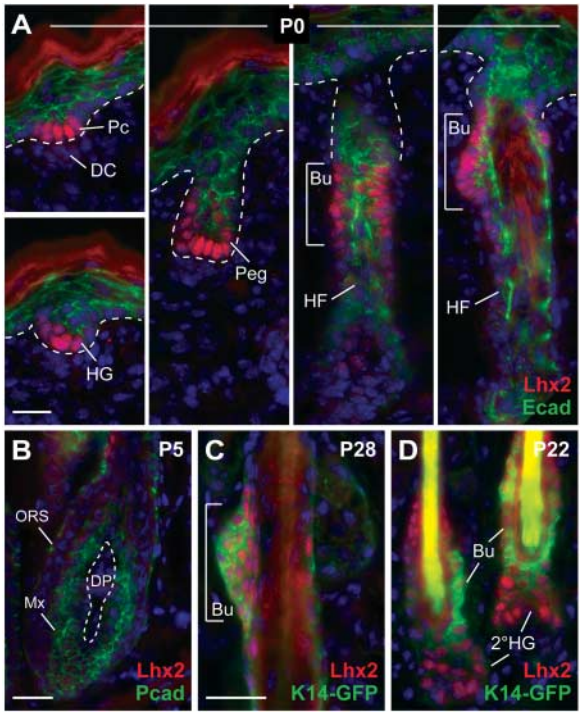



**Fig. 2.** Lhx2 is expressed in early hair progenitors and postnatal stem cells. (**A** to **D**) Back skin sections from mice at indicated ages were stained with antibodies as color-coded and counterstained with DAPI (blue). Lhx2 is expressed in cells at the leading front of invaginating follicles and in the postnatal bulge compartment, but is diminished in mature proliferative hair progenitors (matrix). Abbreviations are as follows: Peg, hair peg; Bu, presumptive bulge; Mx, matrix cells; DC, dermal condensate; DP, dermal papilla; 2°HG, secondary hair germ emerging at the start of the postnatal hair cycle. Scale bars, 20 μm.

Lhx2 was up-regulated 18-fold in the PCAD+ population relative to the PCAD– population, as determined by microarray analysis. Semi-quantitative reverse transcription polymerase chain reaction (RT-PCR) and in situ hybridization confirmed this marked differential expression (fig. S5). When examined by immunofluorescence, Lhx2 first appeared in early hair placodes, and as morphogenesis progressed, became prominent at the leading front of invaginating hair germs and pegs (Fig. 2A). As downgrowth neared completion and hair differentiation began, Lhx2 concentrated in the upper outer root sheath (ORS) at a presumptive site (bulge) of the developing postnatal follicle stem cell compartment (Fig. 2A and fig. S1). Concomitantly, expression diminished at the base of the follicle, where highly proliferative matrix cells give rise to the differentiating inner root sheath and hair shaft (Fig. 2B). In adult follicles, Lhx2 concentrated in the bulge, and as the new hair cycle began, Lhx2 extended to the emerging secondary hair germs (Fig. 2, C and D). Based on these patterns, we posit that Lhx2 functions in specifying the embryonic hair follicle progenitor cells that then persist as bulge stem cells in adult follicles.

To more precisely define Lhx2's role in hair follicle stem cell specification and/or maintenance, we examined its status in various genetic mutant embryos that are defective in different aspects of hair morphogenesis. In the complete absence of hair follicle induction or bulge maintenance, as reflected in β-*catenin* conditionally null [conditional knockout (cKO)] skin, Lhx2 was not expressed (Fig. 3B). In *Shh* knockout embryos, where hair germs are specified but unable to progress, Lhx2 expression was dramatically reduced (Fig. 3C). This positioned Lhx2 downstream of Wnt and Shh, where it could play a role in establishing or expanding the early progenitors necessary for hair follicle morphogenesis.

Bmp signaling is not required for hair follicle induction, even though Bmp ligands and receptors are expressed in embryonic hair germs and in postnatal follicle stem cells. Correspondingly, in *BmpR1a* cKO skin, Lhx2 was expressed in both embryonic hair germs and the presumptive bulge of developing follicles (Fig. 3, D and E). Conversely, Bmp signaling is required for dif-

ferentiation, and in the absence of BmpR1a, proliferating undifferentiated hair progenitor cells accumulate at the follicle base (*27*, *28*). Lhx2 was noticeably enhanced in these follicles, with strong staining throughout the ORS and matrix (Fig. 3, E and F). These cells were also positive for *Shh* and Lef1. Thus, in the absence of terminal hair differentiation, cells accumulating in postnatal *BmpR1a* null follicles resembled early embryonic hair follicle progenitors.

If Lhx2 governs the gene expression program of undifferentiated follicle stem cells or their early progenitors, then misexpression of Lhx2 in interfollicular epidermis might result in an induction of hair follicle progenitor genes. To test this possibility, we engineered *K14-Lhx2* transgenic mice. Although more hair follicles were not induced, Lhx2 markedly suppressed morphological and biochemical signs of epidermal differentiation and failed to produce a functional lipid barrier (Fig. 4A). Most notable was the induction of Tcf3 and Sox9 (Fig. 4B), two key transcription factors of adult hair follicle stem cells (*29*, *30*). Lhx2 also suppressed differentiation in tongue epithelium (fig. S6). These findings suggest that Lhx2 can maintain cells in an undifferentiated state, reinforcing the link between Lhx2 and stemness.

If Lhx2 is required for follicle stem cell maintenance, then its absence could alter the ability of hair follicles to form. In support of this notion, E16 *Lhx2* null embryos displayed an ~40% reduction in the overall density of P-cadherin–positive hair follicles (Fig. 4C), with no noticeable defect in the epidermis or embryo size. A marked reduction in follicle density is a feature of other mouse mutants in key hair follicle morphogenetic genes. Although *Lhx2* KO follicle density was reduced, *Shh*, *Wnt10b*, *Bmp2*, *Bmp4*, and Lef1 expression appeared unaffected in those hair placodes and germs that developed (fig. S7). In *Lhx2* null skin engraftments, follicles appeared morphologically and biochemically indistinguishable from their wild-type counterparts (fig. S8). Taken together, the gain- and loss-of-function studies suggest that Lhx2, reflecting its expression pattern, functions to specify and maintain hair follicle stem cells but does not function in their differentiation.

If Lhx2 maintains the undifferentiated state of embryonic and adult follicle stem cells, then *Lhx2* null follicles might exhibit alterations in the transition of stem cells from the resting (telogen) to the growing (anagen) phase of the postnatal hair cycle. Using skin grafts, we compared the hair cycles of wild-type and *Lhx2* KO follicles. The initial morphogenetic and first postnatal *Lhx2* KO hair cycles progressed similarly to those in the wild type, and by 8 weeks, KO follicles had returned to telogen (fig. S9). By contrast, at 11 weeks, when most wild-type follicles were still in this extended telogen, KO follicles had precociously entered the next hair cycle (Fig. 5A). Moreover, after being shaved at 8 weeks, most wild-type hairs remained in telogen, whereas KO hairs consistently and uniformly grew back within 3 weeks, confirming their shortened resting phase (Fig. 5B).

Immunofluorescence and FACS analyses revealed that KO follicles exhibited diminished

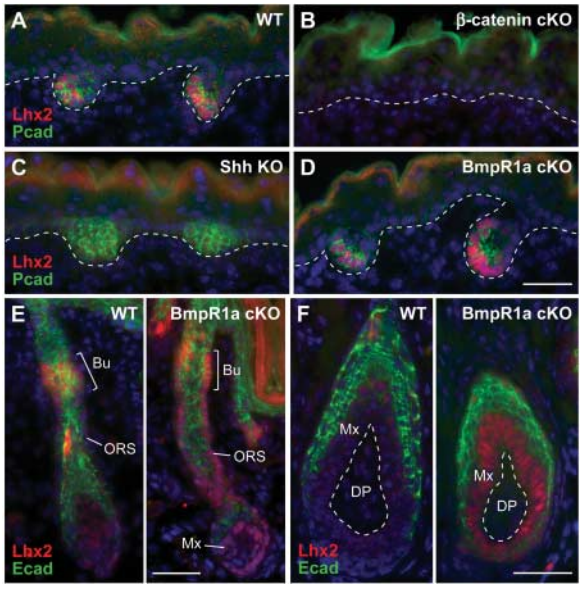



**Fig. 3.** Lhx2 functions downstream of follicle stem cell specification but upstream of their differentiation. (**A**) Lhx2 is expressed in P-cadherin–positive hair germs. Shown is a representative wild-type littermate at E17.5. (**B**) Lhx2 is not expressed in the absence of hair follicle induction, as reflected in the β-*catenin* null skin epithelium (*5*). *K14-Cre* was used to conditionally target β-*catenin*. (**C**) Lhx2 is reduced in *Shh* null hair germs. Follicles are specified in the absence of Shh but fail to progress further (*8*, *10*). (**D**) Lhx2 is expressed in *BmpR1a* conditional null hair germs. (**E** and **F**) In addition to expression in the presumptive bulge, Lhx2 persists in the lower ORS and matrix of neonatal and adult *BmpR1a* null follicles. Unable to undergo terminal differentiation, these *BmpR1a* null, Lhx2-expressing cells appear to be undifferentiated follicle stem cells (*27*, *28*). Scale bars, 40 μm.

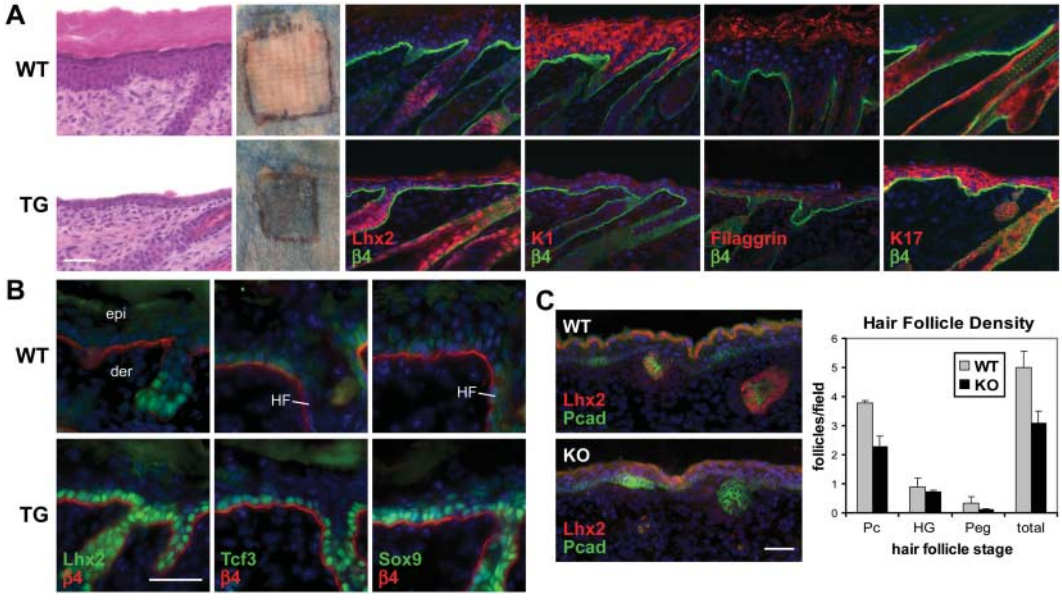



**Fig. 4.** Gain- and loss-of-function studies reveal a role for Lhx2 in promoting follicle stem cell maintenance. (**A**) Transgenic expression of Lhx2 in the interfollicular epidermis suppresses terminal differentiation. Morphological and biochemical features of spinous, granular, and stratum corneum stages of epidermal differentiation are diminished in d8 (day 8 after graft) skin from *K14-Lhx2* transgenic mice (TG) relative to wild-type littermates. Patches outlined in black denote skins grafted onto *Nude* mice and subjected to a β-galactosidase substrate exclusion assay to test for an intact epidermal barrier. The background absorption of dye in the surrounding *Nude* skin arises from defective hair follicle orifices. (**B**) Immunofluorescence reveals an induction of follicle stem cell markers in d4 *K14-Lhx2* epidermis. (**C**) Loss of Lhx2 reduces hair morphogenesis. Representative skin sections from E16 littermates show comparable epidermal differentiation but fewer follicles at all stages in *Lhx2* null embryos. The graph provides quantification from multiple sections of three embryos. Scale bars, 40 μm.

expression of CD34, a surface marker of bulge stem cells (Fig. 5C) (*19*). This reduction in CD34 was observed irrespective of hair cycle number or stage. Other stem cell markers that we examined were comparably expressed in wild-type and KO bulges (fig. S10).

Although CD34 marks adult stem cells, it is not found in embryonic skin progenitors, suggesting that its reduction could be an indication of enhanced proliferative activity within KO follicle stem cells. This hypothesis was supported by bromodeoxyuridine (BrdU) pulse-chase experiments conducted before marked deviations in hair cycling (Fig. 5D). Only the wild-type follicle bulge compartment retained appreciable BrdU label administered at the onset of anagen and chased for 4 weeks (*19*, *31*). By contrast, KO hair follicles displayed very few label-retaining cells (LRCs), as confirmed and quantified by flow cytometry.

The reduction in label retention was accompanied by enhanced proliferation within the KO bulge. After a 4-hour BrdU pulse during full anagen, the percentage of S-phase labeled bulge cells was twice as high as normal (Fig. 5E and fig. S11). By contrast, the number of S-phase cells in the interfollicular epidermis/ORS of wild-type and KO skins was comparable, underscoring the specificity of this hyperproliferation. The elevated proliferative activity of the KO bulge did not appear to alter the overall size of the stem cell niche. We conclude that without Lhx2, follicle stem cells are more readily activated to proliferate and differentiate along the hair lineage. On the other hand, Lhx2 is not sufficient to induce quiescence, because transgenic expression did not suppress proliferation or induce CD34 in the skin epithelium.

Our ability to isolate and transcriptionally profile embryonic hair placodes and interfollicular epidermis has enabled us to substantiate genes previously implicated in hair and epidermal development and to uncover differences that could be important in orchestrating lineage specification of multipotent skin progenitors. Lhx2 has served as a paradigm for testing this premise, and our studies reveal that it functions as a molecular brake in regulating the switch between hair follicle stem cell maintenance and activation. Although follicles can be specified embryonically without Lhx2, their overall numbers are reduced, and *Lhx2* null follicles that do form are not proficient in maintaining the resting state and precociously activate. Once committed, cells no longer require or express Lhx2 and progress along a normal program of terminal differentiation (fig. S12). Finally, Lhx2 is the first identified marker expressed specifically by both embryonic hair placodes and postnatal follicle stem cells of the bulge. Lhx2 now provides a means to dissect the transcriptional mechanisms that underlie stem cell maintenance within the hair follicle.

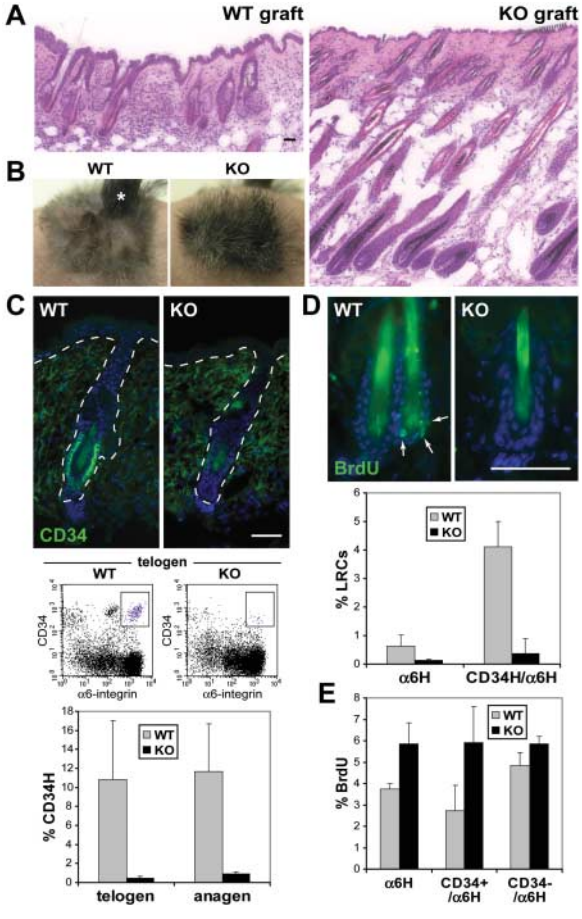



**Fig. 5.** Lhx2 maintains follicle stem cells in a quiescent, inactive state. (**A**) Histology of wild-type and *Lhx2* KO skins at 11 weeks after graft. Wild-type follicles were still in telogen, but KO follicles prematurely entered anagen. (**B**) Upon shaving at 8 weeks, wild-type hairs did not grow back, confirming their resting state. In contrast, KO hairs grew back within 3 weeks. A small portion (*) of the wild-type graft grew hairs because of a wound. (**C**) CD34 expression is dramatically reduced in KO follicle bulges, irrespective of whether they are in telogen or anagen. Shown are representative immunofluorescence images and FACS profiles of telogen follicles at 8 weeks, and CD34 quantification by flow cytometry in telogen and anagen follicles during the first postnatal cycle. (**D**) Loss of BrdU label retention in KO follicles. After a 3-day BrdU pulse on days 26 to 28 at the onset of anagen in both wild-type and KO skin grafts, and a 4-week chase when follicles had entered telogen (fig. S9), LRCs concentrated in the infrequently dividing bulge stem cells of wild-type follicles, but LRCs were diminished in Lhx2 KO skin. Shown are skin sections stained with antibody to BrdU and results of quantification by flow cytometry. (**E**) Increased BrdU incorporation by KO follicle stem cells. After a 4-hour BrdU pulse at day 40, when wild-type and KO follicles were in mid-anagen of their first postnatal hair cycle (fig. S9), cells were isolated and α6-integrin–expressing S-phase cells were quantified by flow cytometry. Scale bars, 40 μm.

### References and Notes


1. M. H. Hardy, *Trends Genet.* **8**, 55 (1992).
2. S. E. Millar, *J. Invest. Dermatol.* **118**, 216 (2002).
3. R. Schmidt-Ullrich, R. Paus, *Bioessays* **27**, 247 (2005).
4. R. DasGupta, E. Fuchs, *Development* **126**, 4557 (1999).
5. J. Huelsken, R. Vogel, B. Erdmann, G. Cotsarelis, W. Birchmeier, *Cell* **105**, 533 (2001).
6. V. A. Botchkarev *et al.*, *Nat. Cell Biol.* **1**, 158 (1999).
7. C. Jamora, R. DasGupta, P. Kocieniewski, E. Fuchs, *Nature* **422**, 317 (2003).
8. C. Chiang *et al.*, *Dev. Biol.* **205**, 1 (1999).
9. A. E. Oro, K. Higgins, *Dev. Biol.* **255**, 238 (2003).
10. B. St-Jacques *et al.*, *Curr. Biol.* **8**, 1058 (1998).
11. B. L. Hogan, *Cell* **96**, 225 (1999).
12. J. Pispa, I. Thesleff, *Dev. Biol.* **262**, 195 (2003).
13. Z. Yue, T. X. Jiang, R. B. Widelitz, C. M. Chuong, *Nature* **438**, 1026 (2005).
14. A. Vaezi, C. Bauer, V. Vasioukhin, E. Fuchs, *Dev. Cell* **3**, 367 (2002).
15. E. Fuchs, S. Raghavan, *Nat. Rev. Genet.* **3**, 199 (2002).
16. J. Tao *et al.*, *Development* **132**, 1021 (2005).
17. S. B. Ting *et al.*, *Science* **308**, 411 (2005).
18. I. Arnold, F. M. Watt, *Curr. Biol.* **11**, 558 (2001).
19. C. Blanpain, W. E. Lowry, A. Geoghegan, L. Polak, E. Fuchs, *Cell* **118**, 635 (2004).
20. R. J. Morris *et al.*, *Nat. Biotechnol.* **22**, 411 (2004).
21. M. Ito *et al.*, *Nat. Med.* **11**, 1351 (2005).
22. V. Levy, C. Lindon, B. D. Harfe, B. A. Morgan, *Dev. Cell* **9**, 855 (2005).
23. F. D. Porter *et al.*, *Development* **124**, 2935 (1997).
24. S. Bulchand, E. A. Grove, F. D. Porter, S. Tole, *Mech. Dev.* **100**, 165 (2001).
25. J. Hirota, P. Mombaerts, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 8751 (2004).
26. P. Pinto do O, K. Richter, L. Carlsson, *Blood* **99**, 3939 (2002).
27. T. Andl *et al.*, *Development* **131**, 2257 (2004).
28. K. Kobielak, H. A. Pasolli, L. Alonso, L. Polak, E. Fuchs, *J. Cell Biol.* **163**, 609 (2003).
29. B. J. Merrill, U. Gat, R. DasGupta, E. Fuchs, *Genes Dev.* **15**, 1688 (2001).
30. V. P. Vidal *et al.*, *Curr. Biol.* **15**, 1340 (2005).
31. G. Taylor, M. S. Lehrer, P. J. Jensen, T. T. Sun, R. M. Lavker, *Cell* **102**, 451 (2000).



32. In the supporting online materials and methods, we cite our many colleagues for their generous contributions of mice and reagents. We acknowledge colleagues in the Fuchs laboratory for constructive discussions and criticisms. We thank N. Stokes, J. Dela Cruz, S. Mazel, T. Shengelia, A. Viale, J. Li, and F. Berguido for invaluable technical assistance. E.F. is a Howard Hughes Medical Institute investigator. This work was supported in part by grant R01-AR050452 (E.F.) from NIH.


### Supporting Online Material

www.sciencemag.org/cgi/content/full/312/5782/1946/DC1
Materials and Methods
Figs. S1 to S12
Tables S1 and S2
References


28 March 2006; accepted 22 May 2006
10.1126/science.1128004

# Structural Basis of RNA-Dependent Recruitment of Glutamine to the Genetic Code

Hiroyuki Oshikane,[1]* Kelly Sheppard,[2]* Shuya Fukai,[1] Yuko Nakamura,[1] Ryuichiro Ishitani,[1] Tomoyuki Numata,[1] R. Lynn Sherrer,[2] Liang Feng,[2] Emmanuelle Schmitt,[3] Michel Panvert,[3] Sylvain Blanquet,[3] Yves Mechulam,[3] Dieter Söll,[2]† Osamu Nureki[1]†

Glutaminyl–transfer RNA (Gln-tRNA$^{Gln}$) in archaea is synthesized in a pretranslational amidation of misacylated Glu-tRNA$^{Gln}$ by the heterodimeric Glu-tRNA$^{Gln}$ amidotransferase GatDE. Here we report the crystal structure of the *Methanothermobacter thermautotrophicus* GatDE complexed to tRNA$^{Gln}$ at 3.15 angstroms resolution. Biochemical analysis of GatDE and of tRNA$^{Gln}$ mutants characterized the catalytic centers for the enzyme's three reactions (glutaminase, kinase, and amidotransferase activity). A 40 angstrom–long channel for ammonia transport connects the active sites in GatD and GatE. tRNA$^{Gln}$ recognition by indirect readout based on shape complementarity of the D loop suggests an early anticodon-independent RNA-based mechanism for adding glutamine to the genetic code.

The formation of at least 20 sets of aminoacyl–transfer RNA (aa-tRNA) is a requirement for protein biosynthesis and its accuracy. In most cases, aa-tRNA synthesis is achieved by direct acylation of tRNA with the cognate amino acid catalyzed by the aminoacyl-tRNA synthetases (aaRSs) (*1*). However, a number of aa-tRNAs are made by pretranslational tRNA-dependent amino acid modification processes. These include selenocysteinyl-tRNA (*2*), cysteinyl-tRNA in methanogenic archaea (*3*), and amide aa-tRNAs (Asn-tRNA and Gln-tRNA) in most prokaryotes (*4*). These pretranslational amino acid modifications represent essential, older pathways of aa-tRNA synthesis (*5*–*8*) and are probably the routes by which these amino acids first entered the genetic code (*2*). Lateral gene transfer of an aaRS may have replaced these indirect pathways in some organisms (*9*). These pathways also provide the sole route for the supply of asparagine in many bacteria (*10*) and of cysteine in methanogens (*3*).

Given the mechanistic unity of the synthesis of most aa-tRNAs, it was surprising to discover the lack of conservation in Gln-tRNA formation (*11*). The eukaryotic cytoplasm uses direct acylation of tRNA$^{Gln}$ by glutaminyl-tRNA synthetase (GlnRS), whereas most bacteria and all archaea use a tRNA-dependent transamidation route (*11*). The first step in this route for Gln-tRNA synthesis is the formation of misacylated Glu-tRNA$^{Gln}$ by a "nondiscriminating" (ND) glutamyl-tRNA synthetase (GluRS), an aaRS that is also responsible for the formation of Glu-tRNA$^{Glu}$ (*12*). This mischarged tRNA is then converted by amidation to the properly charged Gln-tRNA$^{Gln}$ by a Glu-tRNA$^{Gln}$ amidotransferase (*11*, *13*).

Nature has two Glu-tRNA$^{Gln}$ amidotransferases; archaea employ the heterodimeric GatDE enzyme (*11*), whereas bacteria use heterotrimeric GatCAB protein (*13*). Each enzyme has one subunit (GatA or GatD) for ammonia production, yet they are structurally different; GatA is homologous to amidases (*13*) and GatD is similar to a type I L-asparaginase (AnsA) (*11*). The other crucial amidotransferase subunit (GatB or GatE) belongs to the Pet112 protein family (*14*), which consists only of highly related GatB/GatE proteins. They catalyze two reactions: (i) the phosphorylation of the γ-carboxyl group of glutamate in Glu-tRNA$^{Gln}$ by a kinase activity (*15*), and (ii) conversion of the tRNA-bound phosphoglutamate to glutamine by an amidotransferase activity (*15*, *16*) using ammonia generated by GatA/GatD. These amidotransferases may be examples of how amino acid–metabolizing enzymes were recruited during evolution to aa-tRNA formation and thus protein synthesis (*4*).

Recently, the crystal structure of the *Pyrococcus abyssi* GatDE apoenzyme was reported (*17*). It revealed that the GatD and GatE subunits form an $\alpha_2\beta_2$ tetramer, and [based on earlier biochemical experiments (*15*)] residues of the glutaminase catalytic site were suggested to be located in GatD. This left unresolved the structural and biochemical definition of all three active sites and their interconnection, as well as the nature and specificity of binding the tRNA substrate. Here we report the crystal structure of the *Methanothermobacter thermautotrophicus* GatDE complexed to

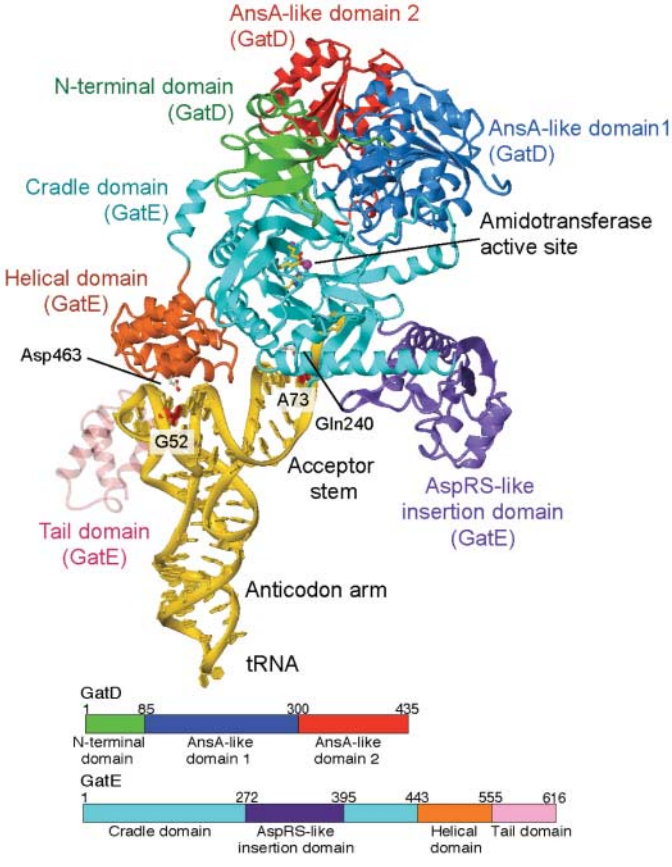



**Fig. 1.** Overall structure of the 1:1 complex of GatDE and tRNA$^{Gln}_1$ (molecules B, D, and F). The protein domains are colored differently; in GatD, the N-terminal domain, AnsA-like domain 1, and AnsA-like domain 2 are shown in light green, blue, and red, respectively; whereas in GatE, the AspRS-like insertion domain, cradle domain, helical domain, and Yqey-like tail domain are colored violet, cyan, orange, and pink, respectively. In the present structure, the Yqey-like tail domain is shown as translucent because its side chains are disordered, despite the fact that the main chain was traced in the electron density map. The bound tRNA molecules are yellow. In GatE, His$^{15}$, Glu$^{157}$, and Glu$^{184}$, which coordinate to an essential $Mg^{2+}$ ion (red), are shown to highlight the Glu-tRNA$^{Gln}$ kinase and amidotransferase sites. Gln$^{240}$, which recognizes A73 (red), and Asp$^{463}$, which recognizes G52 (red), are indicated. All figures of the molecular models were prepared with the program CueMol (www.cuemol.org/).

[1]Department of Biological Information, Graduate School of Bioscience and Biotechnology, Tokyo Institute of Technology, 4259 Nagatsuta-cho, Midori-ku, Yokohama-shi, Kanagawa 226-8501, Japan. [2]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520–8114, USA. [3]Laboratoire de Biochimie, Unité Mixte de Recherche 7654, CNRS-Ecole Polytechnique, F-91128 Palaiseau Cedex, France.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: onureki@bio.titech.ac.jp (O.N.); dieter.soll@yale.edu (D.S.)

tRNA$^{Gln}$ at 3.15 Å resolution. Combined with biochemical data, this structure reveals a mechanism of concerted action of the three enzymatic activities regulated by Glu-tRNA$^{Gln}$. The structural implications of tRNA$^{Gln}$ recognition by indirect RNA readout based on shape complementarity of the D loop suggest an early anticodon-independent RNA-based mechanism for adding glutamine to the genetic code.

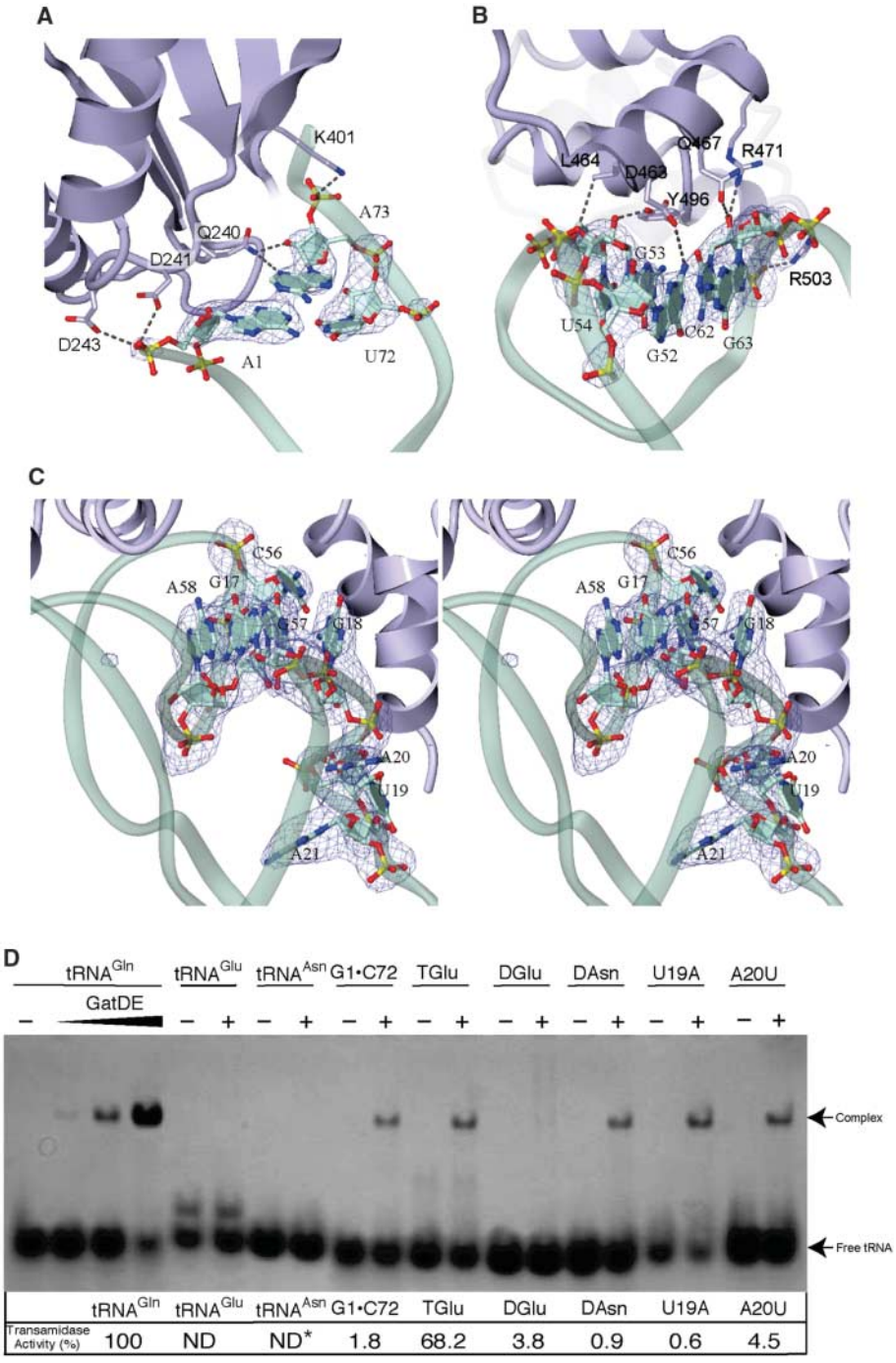



**Fig. 2.** tRNA recognition. (**A**) Recognition of the tRNA acceptor end. (**B**) Recognition of the tRNA TΨC stem. (**C**) Recognition of the tRNA D-loop•TΨC-loop associating region (in stereo). In (A) to (C), $F_o - F_c$ omit maps of the tRNA nucleotides contoured at 5.0, 3.0, and 3.0σ, respectively, are shown. (**D**) Gel retardation assay and transamidase activity of tRNA$^{Gln}_1$ mutants. (Left to right) Wild-type tRNA$^{Gln}_1$ only; wild-type tRNA$^{Gln}_1$ with GatDE (molecular ratios of 1:0.5, 1:2, and 1:8, respectively); tRNA$^{Glu}$ without and with GatDE; tRNA$^{Asn}$ without and with GatDE; tRNA$^{Gln}_{G1•C72}$ without and with GatDE; tRNA$^{Gln}_{TGlu}$ without and with GatDE; tRNA$^{Gln}_{DGlu}$ without and with GatDE; tRNA$^{Gln}_{DAsn}$ without and with GatDE; tRNA$^{Gln}_{U19A}$ without and with GatDE; and tRNA$^{Gln}_{A20U}$ without and with GatDE. Transamidase activities for the various tRNA species were also tested. The percentages shown are relative to the activity using wild-type Glu-tRNA$^{Gln}$. ND means no detectable activity. The asterisk indicates that the substrate for transamidation was Asp-tRNA$^{Asn}$.

As in the *P. abyssi* GatDE apoenzyme (*17*), the GatD subunit consists of three domains: the N-terminal domain, the AnsA-like domain 1, and the AnsA-like domain 2 (Fig. 1). The GatE subunit is made up of a cradle domain, an AspRS-like insertion domain of 123 amino acids, a helical domain, and the C-terminal tail domain (residues 562 to 616) (Fig. 1). This tail domain shares considerable sequence similarity with the C-terminal half of *Bacillus subtilis* YqeY, a protein of unknown function (*18*). Its topology [Protein Data Bank (PDB) identification number 1NG6] could be fitted into the electron density of the GatE tail domain, and the main chain could be traced into a three-helix bundle, with the side chains disordered. The orientation of the GatE helical and tail domains in the tRNA complex is different from that in the apoenzyme structure to allow interaction with tRNA$^{Gln}$ (figs. S1 and S2).

tRNA$^{Gln}$ makes contact only with the GatE subunit, and the tRNA anticodon does not interact with the protein (Fig. 1). This was unexpected, because anticodon recognition, as in aaRSs (*19*), would be sufficient for discrimination between tRNA$^{Gln}$ and tRNA$^{Glu}$. GatE enfolds the acceptor-TΨC helix of the tRNA molecule (Fig. 1), as the CCA-adding RNA polymerase does (*20*, *21*). The helical domain and the C-terminal tail domain form a concave surface, which snugly fits the tRNA elbow region comprising the TΨC arm and the D loop (Fig. 1). The helical domain interacts with the minor groove of the TΨC stem, whereas the C-terminal tail domain invades the TΨC-loop•D-loop assembling region. The helical domain folds into a superhelical architecture that is structurally similar to the body domain of the class-II CCA-adding polymerase (*21*, *22*); these enzymes also recognize the tRNA TΨC arm in a shape-complementary manner (*21*).

GatDE accurately discriminates tRNA$^{Gln}$ from tRNA$^{Glu}$ and tRNA$^{Asn}$, and amidates only the misacylated Glu-tRNA$^{Gln}$ (*11*). In contrast, both Glu-tRNA$^{Gln}$ and Asp-tRNA$^{Asn}$ are substrates for GatCAB (*23*). Because GatB lacks the AspRS-like insertion domain found in GatE, it was suggested that this GatE domain may be an antideterminant against tRNA$^{Asn}$ (*17*). However, in the GatDE•RNA$^{Gln}$ structure, the AspRS-like insertion domain does not contact the tRNA (Fig. 1), and we suggest a different role for this domain (see below).

The tRNA$^{Gln}$ ACCA-terminus (positions 73 to 76) enters the central cleft of the GatE cradle domain. At its entrance, the amide group of Gln$^{240}$ forms a bifurcated hydrogen bond with N3 (3.3 Å) of the adenine ring and the ribose 2′-OH (2.5 Å) of the discriminator base A73 (Fig. 2A). This base is conserved in the *M. thermautotrophicus* tRNA$^{Gln}$, tRNA$^{Glu}$, and tRNA$^{Asn}$ species (fig. S3). The mutation of Gln$^{240}$ to Ala resulted in a 2.1-fold loss of the amidotransferase activity (Fig. 3C). Furthermore, Lys$^{401}$ in the cradle domain and

$Arg^{503}$ in the helical domain provide electrostatic interactions with the phosphate backbones of A73 and C74 of $tRNA^{Gln}$, respectively (Fig. 2A). The replacement of $Arg^{503}$ with Ala reduced the amidotransferase activity by 1.7 fold (Fig. 3C).

The helical domain interacts with the minor groove of the $tRNA^{Gln}$ TΨC stem (Fig. 1). The $Tyr^{496}$ hydroxyl group hydrogen-bonds (2.6 Å) to the 2′-hydroxyl of G53 in the TΨC stem, whereas the δ-amino group of $Gln^{467}$ and the ε-amino group of $Arg^{471}$ hydrogen-bond with the 2′-OH of C62 (Fig. 2B). $Arg^{503}$ electrostatically interacts (3.6 Å) with the backbone phosphate of C62 (Fig. 2B). Base-specific recognition is observed for G52, whose 2-amino group hydrogen-bonds to the β-carboxyl group of $Asp^{463}$ (Fig. 2B). Changing this residue to Ala reduced amidotransferase activity by 1.5 fold, and replacement of $Arg^{503}$ with Ala lowered the transamidation activity by 1.7 fold (Fig. 3C). However, G52, G53, and C62 are common to the $tRNA^{Gln}$, $tRNA^{Glu}$, and $tRNA^{Asn}$ species (fig. S3). Thus, the GatDE interactions with the TΨC stem do not contribute to tRNA discrimination but may anchor the TΨC arm to properly position the TΨC-loop•D-loop–assembling region toward the tail domain.

The two helices in the tail domain form a concave surface to accommodate the TΨC and D loops (Fig. 1). Although the side chains are disordered, two helices of the tail domain closely contact the TΨC-loop•D-loop–assembling region, especially C56 and G18 (Fig. 2C and fig. S4). As a result, the Watson-Crick type tertiary base pair between C56 and G18 is significantly twisted (Fig. 2C). The concave surface formed by the two helices prevents the accommodation of larger TΨC and D loops; this appears to be crucial for the discrimination of $tRNA^{Gln}$ from $tRNA^{Glu}$. Comparison of the relevant tRNA sequences (fig. S3) reveals that $tRNA^{Glu}$ possesses a D loop that is three nucleotides larger than those of both $tRNA^{Gln}$ species. This bigger loop may act as an antideterminant in $tRNA^{Glu}$ recognition. On the other hand, $tRNA^{Asn}$ has a D loop of different nucleotide sequence yet similar in size to that of $tRNA^{Gln}$.

To experimentally determine the basis for $tRNA^{Gln}$ recognition by *M. thermautotrophicus* GatDE, we constructed tRNA mutants [see the supporting online material (SOM)] and (i) tested tRNA recognition (RNA binding) by gel retardation and (ii) determined the tRNA identity (*19*) by assessing the ability of the homologous mutant tRNAs to be enzyme substrates in the amidotransferase reaction. The two methods may produce some differences, because (i) measures RNA binding; whereas for (ii), productive binding is required. Because the structure revealed binding by shape complementarity of the D loop, we prepared mutants of $tRNA^{Gln}$ in which its D loop or TΨC loop was replaced with that of $tRNA^{Glu}$ and the D loop/stem of $tRNA^{Gln}$ was replaced with that of $tRNA^{Asn}$; this resulted in the D-loop chimeras $tRNA^{Gln}_{DGlu}$ and $tRNA^{Gln}_{DAsn}$ and the T-loop chimera $tRNA^{Gln}_{TGlu}$. Although there was slightly reduced binding activity by the $tRNA^{Gln}_{TGlu}$ and $tRNA^{Gln}_{DAsn}$ mutants, $tRNA^{Gln}_{DGlu}$ did not bind (Fig. 2D). In contrast to the gel-shift results, the Glu-$tRNA^{Gln}_{DGlu}$ and Glu-$tRNA^{Gln}_{DAsn}$ species could not be transamidated by GatDE, whereas Glu-$tRNA^{Gln}_{TGlu}$ supported the transamidation reaction (Fig. 2D). Similarly, the glutamylated U19A and A20U $tRNA^{Gln}_{1}$ mutants did not support amidotransferase activity, although these mutant tRNAs bind to GatDE (Fig. 2D). Thus, the C-terminal helical and tail domains of GatE play a key role in discriminating $tRNA^{Gln}$ from $tRNA^{Glu}$ by shape complementarity as well as by specific nucleotide recognition.

Another significant identity element difference between $tRNA^{Gln}$, $tRNA^{Glu}$, and $tRNA^{Asn}$ might be the first base pair of the acceptor helix; this is A1•U72 in $tRNA^{Gln}$ and G1•C72 in $tRNA^{Glu}$ and $tRNA^{Asn}$ (fig. S3). The G1•C72 $tRNA^{Gln}$ mutant showed 50 times less amidotransferase activity, with the GatE-binding

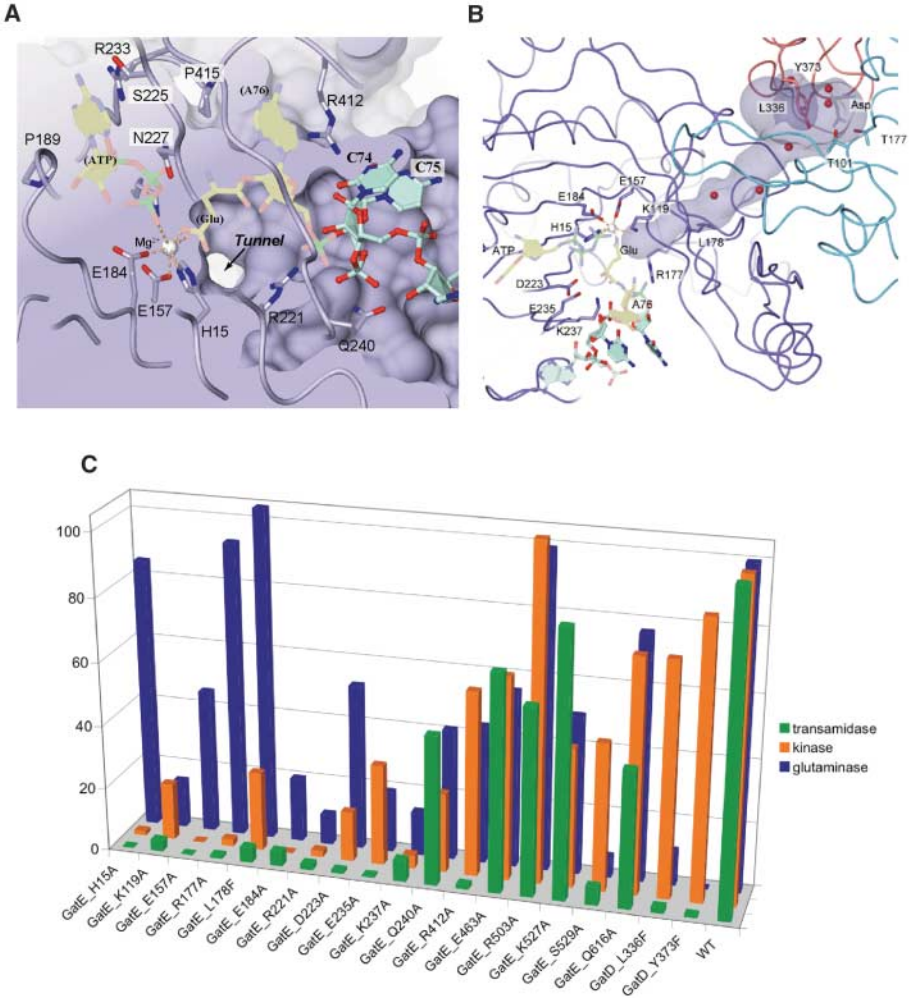



**Fig. 3.** Catalysis by GatDE. (**A**) The Glu-$tRNA^{Gln}$ kinase and amidotransferase center of the GatE subunit. The disordered A76, the attached glutamyl moiety, and substrate ATP are modeled in yellow. The catalytically essential $Mg^{2+}$ ion is represented by a silver ball. Beside the $Mg^{2+}$ ion, a molecular tunnel penetrates into the GatD glutaminase site. C74 and C75 of the extended tRNA terminus are shown in cyan. Amino acid residues involved in substrate recognition and catalysis are indicated. (**B**) Molecular tunnel of GatDE. The molecular tunnel connects the GatD glutaminase site and the GatE kinase/amidotransferase site. GatE is represented by a violet wire model; two GatD molecules are represented by cyan and orange wire models. In the GatD glutaminase site, the catalytic $Thr^{101}$ and $Thr^{177}$ residues are indicated, as well as a modeled L-aspartate, a reaction product, which was reported to bind in the apo form of GatDE from *Pyrococcus abyssi* (*17*). In the GatE active site, the modeled A76 attached with glutamate and substrate ATP are shown in yellow. C74 and C75 of the tRNA terminus are indicated in cyan. The observed water molecules, or possibly ammonia molecules, are indicated by red balls with a diameter of 0.5 Å. (**C**) The glutaminase, Glu-$tRNA^{Gln}$ kinase, and amidotransferase activities (colored blue, orange, and green, respectively) of 17 GatDE mutants. The saturated activities relative to those of the wild-type GatDE are shown. The activity measurements were performed three to five times, and the experimental values were averaged.

activity unaffected (Fig. 2D). Therefore, the G1•C72 base pair in $tRNA^{Asn}$ will be a major antideterminant for GatDE recognition.

The $tRNA^{Gln}$ ACCA-terminus enters the central cavity of the GatE cradle domain; the bases of C74 and C75 stack on each other, and the O2 atom of C74 hydrogen-bonds to the $Arg^{412}$ ε-amino group (Fig. 3A). In contrast to C74 and C75, we did not observe electron density corresponding to A76. Possibly glutamylation is a prerequisite to place A76 in its binding site. In the vicinity of C75 there is a pocket large enough to accommodate glutamyl adenosine.

In the catalytic pocket, $Mg^{2+}$ is observed; it is coordinated with the strictly conserved $His^{15}$, $Glu^{157}$, and $Glu^{184}$ residues. This pocket is expected to house the Glu-$tRNA^{Gln}$ kinase and amidotransferase activities, and the coordinated $Mg^{2+}$ should play a catalytic role in both reactions. By molecular dynamics, we then modeled glutamyl adenosine (position 76) in the catalytic pocket (Fig. 3A), so that the adenine ring was sandwiched between $Arg^{412}$ and $Pro^{415}$ (Fig. 3A), the N6 atom hydrogen bonded with the main-chain carbonyl group of $Val^{413}$, and the Oε atom of the glutamyl moiety (covalently attached to the 2′-OH of A76) was pointed toward the $Mg^{2+}$ in order to be ready for phosphorylation. The α-carbonyl and α-amino groups of the glutamyl moiety are recognized by the conserved amino acids $Lys^{237}$ and $Glu^{235}$, respectively, through hydrogen-bonding interactions (Fig. 3A). The phosphate group of A76 forms a hydrogen bond with $Arg^{221}$ in the GatE cradle domain, which simultaneously interacts electrostatically with the phosphate group of C75 (Fig. 3A). The conserved $Arg^{177}$ also provides an electrostatic interaction with the A76 phosphate moiety (Fig. 3B).

Before the amidotransferase reaction, the γ-carboxyl group of the glutamyl moiety attached to A76 is phosphorylated with concurrent hydrolysis of adenosine triphosphate (ATP) to adenosine diphosphate (ADP). We therefore constructed a docking model of GatDE and ATP by superposing the ADP/$AlF_x$-bound GatB structure (*24*) onto that of GatE (Fig. 3A). This docking model is reliable because the ATP-binding pocket exhibits 53.6% amino acid identity and root mean square deviation of 0.665 Å (28 Cα atoms) between GatE and GatB, with the ATP-interacting $Pro^{189}$, $Ser^{225}$, $Asn^{227}$, and $Arg^{233}$ being strictly conserved. In this model, the γ-phosphate group of ATP coordinates to the essential $Mg^{2+}$ (Fig. 3A). Thus, $His^{15}$, $Glu^{157}$, $Glu^{184}$, the γ-carboxyl group of the glutamate attached to the tRNA, the γ-phosphate of the ATP substrate, and a water molecule octahedrally coordinate with the $Mg^{2+}$ ion (Fig. 3A). This $Mg^{2+}$ coordination may accelerate the polarization of the γ-phosphorus group of ATP and the γ-carboxyl group of the tRNA-bound glutamate, promoting the nucleophilic reaction that forms γ-phosphoryl-glutamyl-$tRNA^{Gln}$.

To corroborate the assignment of the three catalytic sites by functional enzymatic tests (glutaminase, kinase, and transamidase activities), mutant GatDE enzymes were generated in which critical amino acids were replaced by alanine (see SOM). Earlier studies (*15*) showed the essential nature of the GatD residues $Thr^{101}$, $Thr^{177}$, $Asn^{178}$, and $Lys^{254}$ for glutaminase activity of the heterodimeric enzyme. However, the mutant enzymes retained kinase and amidotransferase activity when ammonium chloride was the amide donor (*15*). Analysis of the GatE structure assigned the kinase and transamidase active sites to the vicinity of the $Mg^{2+}$ with the strictly conserved residues $His^{15}$, $Glu^{157}$, and $Glu^{184}$ (Fig. 3A). The enzymes resulting from mutation of these amino acids lost the ATPase and amidotransferase activities but kept the glutaminase function (Fig. 3C). A similar pattern of activities was found (Fig. 3C) in mutant enzymes with alterations in other assigned conserved amino acids [$Arg^{221}$, $Lys^{237}$, and $Arg^{412}$ (Fig. 3A) and $Arg^{177}$, $Asp^{223}$, and $Glu^{235}$ (Fig. 3B)]. The reduction of glutaminase activity in the mutant $Arg^{221}$Ala, $Glu^{235}$Ala, and $Lys^{237}$Ala enzymes supports the observed coupling of the GatD glutaminase activity with Glu-$tRNA^{Gln}$ binding by GatE (*15*).

The GatDE•$tRNA^{Gln}$ complex structure makes clear that the kinase/amidotransferase catalytic site on GatE is located 40 Å apart from the glutaminase site on GatD (Fig. 1). How does ammonia generated at the latter site move to the kinase/amidotransferase site on the other subunit? In the vicinity of the catalytic $Mg^{2+}$ ion, there is a hole (Fig. 3A) that is the exit of an intersubunit 40 Å–long molecular tunnel, which connects the GatD glutaminase site and the GatE kinase/amidotransferase site (Fig. 3B). This tunnel penetrates one GatE and two GatD subunits, the entrance being located on the interface of the GatD dimer (Fig. 3B). With an average diameter of 2 Å, the tunnel should be capable of transporting ammonia. The amino acids that line the inside of the tunnel have hydrophilic properties; a sequence alignment reveals that they are strictly conserved in archaea. Therefore, we propose that the ammonium generated at the glutaminase site in GatD is transported along the molecular tunnel to the kinase center of GatE, where it amidates γ-phosphoryl-Glu-$tRNA^{Gln}$. In the present structure, we observed several electron densities, corresponding to water or possibly ammonium molecules along the tunnel (Fig. 3B), which suggests that the mechanism of ammonia transport resembles that of potassium transport by a $K^+$ channel (*25*). Mutations of two amino acids [$Leu^{178}$(GatE)Phe and $Leu^{336}$(GatD)Phe] located along the molecular tunnel (Fig. 3B) significantly reduced the amidotransferase activity (Fig. 3C). The $Tyr^{373}$(GatD)Phe mutation, situated at the entrance of the tunnel (Fig. 3B), reduced the amidotransferase and glutaminase activities considerably (Fig. 3C); this suggests that $Tyr^{373}$ plays a role in withdrawing and attracting the generated ammonia into the tunnel. Which residue(s) deprotonate the transported ammonium at the exit of the tunnel and direct it to react with the phosphorylated glutamate carboxyl group? Sequence and functional analyses indicated that $Lys^{119}$, whose ε-amino group is activated by the adjacent GatE residues $Gln^{131}$, $Glu^{157}$, and $Glu^{184}$, and by GatD $Tyr^{55}$, is involved in the deprotonation of the ammonium (Fig. 3, B and C).

Generally, enzymes using glutamine as an amide donor are thought to have a molecular tunnel, through which ammonium is transferred from the glutaminase site to the second synthetase site (*16*, *26*). A recent structural analysis reported that glucosamine-6-phosphate synthase opens its molecular tunnel by rotating the indole ring of a Trp

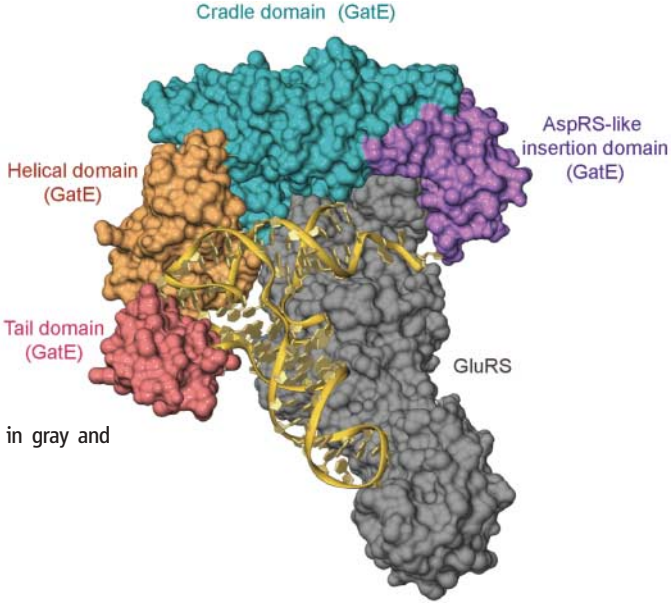



**Fig. 4.** Ternary complex formation between GatE, GluRS, and $tRNA^{Gln}$. Docking of *T. thermophilus* GluRS complexed with $tRNA^{Glu}$ (*29*) onto the present GatE•$tRNA^{Gln}$ complex was accomplished by superposing the complexed tRNA structures. In GatDE, the AspRS-like insertion domain, cradle domain, helical domain, and tail domain are colored violet, cyan, orange, and pink, respectively. GluRS and tRNA are shown in gray and yellow, respectively.

residue upon the binding of the glutamine substrate to the glutaminase site (*27*). Elucidation of a similar tunnel-gating mechanism will await the structure determination of GatDE complexed with acylated Glu-tRNA$^{Gln}$.

The presence of misacylated Glu-tRNA$^{Gln}$ as the required precursor for Gln-tRNA$^{Gln}$ prompted the question of whether there is a complex between the ND-GluRS and the Glu-tRNA$^{Gln}$ amidotransferase, where the misacylated tRNA would be transferred from the synthetase to the amidotransferase by substrate channeling (*28*). This would ameliorate the task of EF-Tu in keeping protein synthesis accurate by not transporting Glu-tRNA$^{Gln}$ to the ribosome (*29*). Preliminary experiments suggest the existence of a complex between *M. thermautotrophicus* GluRS and GatDE; we then docked *T. thermophilus* GluRS complexed with tRNA$^{Glu}$ (*30*) onto the present GatDE•tRNA$^{Gln}$ complex by superposing the tRNA structures (Fig. 4). GluRS fits well into the concave surface formed by the GatE cradle and the AspRS-like insertion domains, and the bound tRNA$^{Gln}$ extensively interacts with the AspRS-like insertion domain with only minor steric clashes. A similar docking of the AspRS•tRNA$^{Asp}$ complex (*31*) did not succeed because of serious steric hindrance, especially with the AspRS-like insertion domain. Thus, the AspRS-like insertion domain may facilitate formation of a GluRS•GatDE complex while preventing association with AspRS. As a result, the ternary complex would produce and channel only Glu-tRNA$^{Gln}$, which is the sole substrate for transamidation by GatDE (*11*). The presence of such a complex could also account for the lack of the AspRS-like insertion domain in GatCAB, which converts not only Glu-tRNA$^{Gln}$ but also Asp-tRNA$^{Asn}$ to the corresponding amide aa-tRNAs (*23*).

The putative GluRS•GatDE complex, in which tRNA$^{Gln}$ interacts simultaneously with the ND-GluRS and the AspRS-like GatE insertion domain, mimics one of the aaRS class I/class II/tRNA complexes that have been proposed as intermediates in the evolution of the aaRS classes (*32*). This had further experimental support in the observed ternary complex of the two classes of *Methanosarcina barkeri* lysyl-tRNA synthetase (LysRS) and tRNA$^{Pyl}$ (*33*) and of LysRS1•LysRS2•tRNA$^{other}$ from *Bacillus cereus* (*34*). The presence of such a ternary complex in transamidation would give credence to the idea that this pathway is a remnant of early code formation. Furthermore, unlike temporary aaRSs, GatE mainly uses indirect sequence-independent readout for the recognition of tRNA$^{Gln}$, a mechanism that also operates in EF-Tu (*35*) and tRNA-modifying enzymes (*36*). Taken together, the data suggest that a GatDE-like structure and activity were appended to an existing GluRS to provide the physical and mechanistic means of adding glutamine to the genetic code.


**References and Notes**

1. M. Ibba, D. Söll, *Annu. Rev. Biochem.* **69**, 617 (2000).
2. A. Böck, M. Thanbichler, M. Rother, A. Resch, in *Aminoacyl-tRNA Synthetases*, M. Ibba, C. S. Francklyn, S. Cusack, Eds. (Landes Bioscience, Georgetown, TX, 2004).
3. A. Sauerwald *et al.*, *Science* **307**, 1969 (2005).
4. M. Ibba, D. Söll, *Genes Dev.* **18**, 731 (2004).
5. M. Ibba, H. D. Becker, C. Stathopoulos, D. L. Tumbula, D. Söll, *Trends Biochem. Sci.* **25**, 311 (2000).
6. M. Di Giulio, *J. Mol. Evol.* **55**, 616 (2002).
7. C. R. Woese, G. Olsen, M. Ibba, D. Söll, *Microbiol. Mol. Biol. Rev.* **64**, 202 (2000).
8. P. O'Donoghue, A. Sethi, C. R. Woese, Z. A. Luthey-Schulten, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 19003 (2005).
9. V. Lamour *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **91**, 8670 (1994).
10. B. Min *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 2678 (2002).
11. D. L. Tumbula, H. D. Becker, W. Chang, D. Söll, *Nature* **407**, 106 (2000).
12. J. Lapointe, L. Duplain, M. Proulx, *J. Bacteriol.* **165**, 88 (1986).
13. A. W. Curnow *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 11819 (1997).
14. J. J. Mulero, J. K. Rosenthal, T. D. Fox, *Curr. Genet.* **25**, 299 (1994).
15. L. Feng, K. Sheppard, D. L. Tumbula-Hansen, D. Söll, *J. Biol. Chem.* **280**, 8150 (2005).
16. H. Zalkin, J. L. Smith, *Adv. Enzymol. Related Areas Mol. Biol.* **72**, 87 (1998).
17. E. Schmitt, M. Panvert, S. Blanquet, Y. Mechulam, *Structure* **13**, 1421 (2005).
18. R. A. Laskowski, J. D. Watson, J. M. Thornton, *Nucleic Acids Res.* **33**, W89 (2005).
19. R. Giege, M. Sissler, C. Florentz, *Nucleic Acids Res.* **26**, 5017 (1998).
20. Y. Xiong, T. A. Steitz, *Nature* **430**, 640 (2004).
21. K. Tomita *et al.*, *Nature* **430**, 700 (2004).
22. F. Li *et al.*, *Cell* **111**, 815 (2002).
23. G. Raczniak, H. D. Becker, B. Min, D. Söll, *J. Biol. Chem.* **276**, 45862 (2001).
24. I. Tanaka, personal communication.
25. J. H. Morais-Cabral, Y. Zhou, R. MacKinnon, *Nature* **414**, 37 (2001).
26. F. M. Raushel, J. B. Thoden, H. M. Holden, *Acc. Chem. Res.* **36**, 539 (2003).
27. S. Mouilleron, M. A. Badet-Denisot, B. Golinelli-Pimpaneau, *J. Biol. Chem.* **281**, 4404 (2006).
28. A. Schön, C. G. Kannangara, S. Gough, D. Söll, *Nature* **331**, 187 (1988).
29. F. J. LaRiviere, A. D. Wolfson, O. C. Uhlenbeck, *Science* **294**, 165 (2001).
30. S. Sekine *et al.*, *Nat. Struct. Biol.* **8**, 203 (2001).
31. S. Eiler, A. Dock-Bregeon, L. Moulinier, J. C. Thierry, D. Moras, *EMBO J.* **18**, 6532 (1999).
32. L. Ribas de Pouplana, P. Schimmel, *Cell* **104**, 191 (2001).
33. C. Polycarpo *et al.*, *Mol. Cell* **12**, 287 (2003).
34. S. F. Ataide, B. C. Jester, K. M. Devine, M. Ibba, *EMBO Rep.* **6**, 742 (2005).
35. H. Asahara, O. C. Uhlenbeck, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 3499 (2002).
36. R. Ishitani *et al.*, *Cell* **113**, 383 (2003).


37. We thank M. Kawamoto and N. Shimizu (Japan Synchrotron Radiation Research Institute) for their help in data collection at SPring-8. We are indebted to A. Nakamura, M. Yao, S. Chimnaronk, N. Sakai, and I. Tanaka for sharing atomic coordinates before publication. E.S., M.P., S.B., and Y.M. contributed atomic coordinates before publication. R.L.S. is the recipient of a Kirschstein–National Research Service Award postdoctoral fellowship from the National Institute of General Medical Sciences. This work was supported by a Precursor Research for Embryonic Science and Technology Program grant from Japan Science and Technology to O.N.; by a grant for the National Project on Protein Structural and Functional Analyses from the Ministry of Education, Culture, Sports, Science and Technology (MEXT) to O.N.; by grants from MEXT to O.N. and S.F.; by the Mitsubishi Foundation and Kurata Memorial Hitachi Science and Technology Foundation grants to O.N.; by the National Institute of General Medical Sciences (to D.S.); and by the U.S. Department of Energy (to D.S.). The atomic coordinates and structural factors have been deposited in the PDB (www.rcsb.org; PDB identification code 2D6F).


**Supporting Online Material**

www.sciencemag.org/cgi/content/full/312/5782/1950/DC1
Materials and Methods
SOM Text
Figs. S1 to S5
Table S1
References

7 April 2006; accepted 26 May 2006
10.1126/science.1128470

# Ammonia Channel Couples Glutaminase with Transamidase Reactions in GatCAB

**Akiyoshi Nakamura,[1]* Min Yao,[1]* Sarin Chimnaronk,[1,2] Naoki Sakai,[1] Isao Tanaka[1]†**

**The formation of glutaminyl transfer RNA (Gln-tRNA$^{Gln}$) differs among the three domains of life. Most bacteria employ an indirect pathway to produce Gln-tRNA$^{Gln}$ by a heterotrimeric glutamine amidotransferase CAB (GatCAB) that acts on the misacylated Glu-tRNA$^{Gln}$. Here, we describe a series of crystal structures of intact GatCAB from *Staphylococcus aureus* in the apo form and in the complexes with glutamine, asparagine, $Mn^{2+}$, and adenosine triphosphate analog. Two identified catalytic centers for the glutaminase and transamidase reactions are markedly distant but connected by a hydrophilic ammonia channel 30 Å in length. Further, we show that the first U-A base pair in the acceptor stem and the D loop of tRNA$^{Gln}$ serve as identity elements essential for discrimination by GatCAB and propose a complete model for the overall concerted reactions to synthesize Gln-tRNA$^{Gln}$.**

Accurate translation of the genetic code into proteins requires high efficiency and fidelity of tRNA-aminoacylation (*1*). This vital process is governed by a fundamental family of enzymes called aminoacyl-tRNA synthetases (aaRS), which catalyze the direct attachment of amino acids to the 3′-ends of the corresponding tRNAs. However, Gln-tRNA$^{Gln}$ is synthesized by two different pathways in nature (*2–4*). In the eukaryotic cy-

toplasm and in some bacteria, glutamine is ligated directly to $tRNA^{Gln}$ by glutaminyl-tRNA synthetase (GlnRS), whereas a noncanonical indirect pathway is employed in the majority of bacteria and all archaea, which do not possess GlnRS, through a two-step process (*2*, *5*). During the first step, $tRNA^{Gln}$ is misacylated with glutamic acid by a nondiscriminating glutamyl-tRNA synthetase (GluRS), which recognizes both $tRNA^{Glu}$ and $tRNA^{Gln}$. Thereafter, the mischarged Glu-$tRNA^{Gln}$ is transformed into Gln-$tRNA^{Gln}$ by Glu-$tRNA^{Gln}$–dependent amidotransferase (Glu-AdT). Glu-AdT converts Glu-$tRNA^{Gln}$ into Gln-$tRNA^{Gln}$ by initially activating Glu-$tRNA^{Gln}$ into γ-phosphoryl-Glu-$tRNA^{Gln}$ at the expense of ATP, which is subsequently transamidated into Gln-$tRNA^{Gln}$ using ammonia generated by hydrolysis of glutamine or asparagine. Moreover, Asn-$tRNA^{Asn}$ can also be synthesized through such an indirect pathway by Asp-$tRNA^{Asn}$–dependent amidotransferase (Asp-AdT) in organisms lacking asparaginyl-tRNA synthetase (*6*, *7*).

Bacterial Glu-AdTs are heterotrimeric proteins composed of A, B, and C subunits, and have therefore been named glutamine amidotransferase CAB (GatCAB), while archaea employ the heterodimeric GatDE enzymes (*8*, *9*). Although bacterial GatCAB acts as both Glu- and Asp-AdT, GatDE specifically recognizes only Glu-$tRNA^{Gln}$ (*9*). It is intriguing that the three domains in nature use distinct enzymes to synthesize Gln-$tRNA^{Gln}$ in protein synthesis. Based on the primary sequences analysis, GatA is homologous to the amidase enzymes, whereas GatD is closer to the L-asparaginases (*9*, *10*). Therefore, GatA and GatD are structurally different but are assumed to play the same role as the glutaminase subunit for ammonia production. On the other hand, GatB is highly related exclusively to its archaeal counterpart, GatE (*9*, *10*). GatE alone can produce γ-phosphoryl-Glu-$tRNA^{Gln}$, which suggests that the activation of Glu-$tRNA^{Gln}$ is achieved only by the GatB or GatE subunits (*11*). Notably, the glutaminase activity and the activation of Glu-$tRNA^{Gln}$ into γ-phosphoryl-Glu-$tRNA^{Gln}$ are tightly coupled upon the binding of Glu-$tRNA^{Gln}$ (*11*, *12*). Hence, these modules of distinct functional enzymes were assembled to form a more sophisticated protein complex and acquired novel mechanisms to expand the genetic code during evolution.

Recently, the $\alpha_2\beta_2$ tetramer structure of archaeal GatDE from *Pyrococcus abyssi* was reported at 3.0 Å resolution (*10*); however, the precise mechanisms for each of the overall reactions coupling the glutaminase with the kinase and transamidase activities, regulated by Glu-$tRNA^{Gln}$, remain elusive. To address this issue, we initiated the crystal structure analysis of bacterial GatCAB from the pathogenic bacterium, *S. aureus Mu50* strain. The crystal structure of apo GatCAB was solved at a resolution of 2.5 Å, by the single-wavelength anomalous diffraction (SAD) method, and refined to a final *R*-factor of 23.8% (*R*-free = 27.5%) using the LAFIRE program running with CNS (*13*–*15*) (Fig. 1A). The asymmetric unit contained one GatCAB heterotrimeric molecule, in which the flexible C-terminal region of GatB (residues 412 to 475) was disordered and could not be modeled. The structure immediately highlights an important role of the C subunit as a stabilizer of the protein complex. GatC is featured by an unstructured extended loop flanked by two helices at its N terminus and two β strands at its C terminus (Fig. 1B). GatC wraps around the interface region as a belt and makes extensive interactions with both GatA and GatB. The amphipathic Cα1 and Cα2 helices form a helical bundle with the hydrophobic core of GatA (Aα9-11, 13), thus relieving the local hydrophobicity (Fig. 1C). The internal loop region of GatC crosses over the loop-rich side of GatB, stabilized through a hydrogen bond network by two invariant $Arg^{64}$ and $Asp^{66}$ residues of GatC (Fig. 1D). The association of GatAB complex is further assured by the hydrophobic interactions of Cα3 helix of GatC with GatAB. Interestingly, the C-terminal Cβ1-2 strands make an antiparallel sheet with a β hairpin (Bβ8-9) of GatB, forming a four-stranded sheet, resembling the corresponding region in GatDE (Fig. 1E).

GatA is a single-domain protein consisting of a central, mixed 11-stranded β sheet core covered by double layers of α helices on the top and bottom (fig. S2). The highest structural similarity was found in malonamidase E2 (MAE2) from *Bradyrhizobium japonicum*

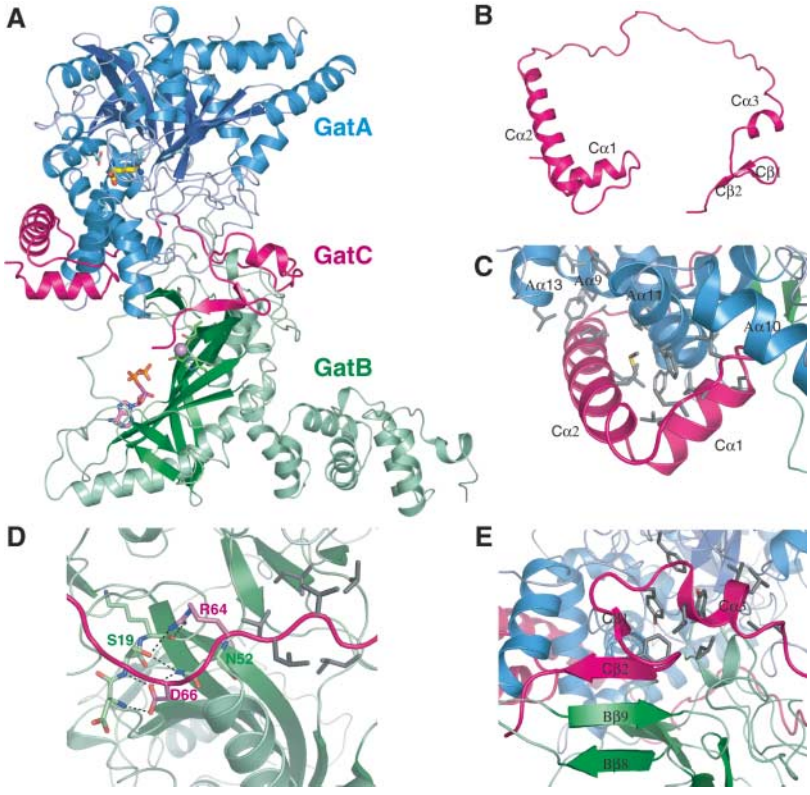



**Fig. 1.** Bacterial GatCAB complex fastens a molecular belt. (**A**) Front-view ribbon diagram of the overall structure of *S. aureus* GatCAB/glutamine complex at 2.3 Å resolution, depicted in three different colors for each subunit: blue, green, and magenta for GatA, GatB, and GatC, respectively. Glutamine in the active site of GatA is drawn as yellow stick representations, whereas the purple sphere is the magnesium ion found in the active site of GatB. ADP from the cocrystal structure with ADP–$AlF_4^-$ is shown together as pink sticks. This color code is used throughout all the figures. (**B**) Top view of annularly shaped GatC. (**C**) Amphipathic helices at the N terminus of GatC form a helical bundle with the hydrophobic core of GatA (hydrophobic residues are colored gray). (**D**) Detailed interactions between the internal loop region of GatC and GatAB complex. Conserved residues involved are labeled, with hydrogen bonds indicated (≤3.2 Å, dashed black line). Polar interactions are prominent on the C-terminal side, whereas the hydrophobic interactions are clustered on the opposite side. (**E**) The C terminus of GatC tightens GatAB complex, constructing an antiparallel β sheet and a hydrophobic platform.

[1]Faculty of Advanced Life Sciences, Hokkaido University, Sapporo 060-0810, Japan. [2]Institute of Molecular Biology and Genetics, Mahidol University, Salaya Campus, Nakornpathom 73170, Thailand.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: tanaka@castor.sci.hokudai.ac.jp

[Protein Data Bank (PDB) code 1OCK (*16*)], with a root mean square deviation of 2.1 Å for 405 pairs of Cα atoms compared. GatA faces its loop-rich plane to, and caps a loop-rich side of, the cradle domain of GatB, burying 7% (2680 Å$^2$) of the total surface area in the intersubunit interface (Fig. 1A). There is a putative ammonia channel (described below) running through the middle of the interface, which is surrounded by two layers of interfacial interactions (fig. S1). The inner layer encircling the channel is composed of polar residues, whereas the outer layer is mostly hydrophobic; therefore, GatA is moderately attached to GatB through both hydrogen-bonding and hydrophobic interactions.

Seeking the molecular mechanism of the glutaminase reaction in GatA, we soaked GatCAB crystals in 1 mM glutamine and determined the cocrystal structure by molecular replacement. The glutamine was found in the center of GatA constructed by the amidase signature sequence (fig. S2). The amide group of glutamine is recognized by Asp$^{425}$, and the carboxyl group is recognized by Arg$^{358}$ (Fig. 2A). The side chain of the glutamine is located close to the conserved Ser–*cis*-Ser–Lys catalytic scissors: Ser$^{178}$, Ser$^{154}$, and Lys$^{79}$. Consistent with MAE2 (*16*), in GatA, Ser$^{154}$ is located in the loop enriched in small residues (glycine, serine, and alanine) and forms an unusual cis conformation, which enables itself to deprotonate Ser$^{178}$. Surprisingly, Ser$^{178}$ made a tight covalent bond (1.44 Å) with the amide carbonyl carbon atom of the glutaminyl side chain, revealing the tetrahedral covalent intermediate, which is stabilized by the oxyanion hole constructed by the backbone nitrogen atoms of Thr$^{175}$, Gly$^{176}$, Gly$^{177}$, and Ser$^{178}$ (Fig. 2A). We further observed a continuous electron density sprouting out from the amide group of the side chain of glutamine, which may correspond to amide ammonia liberated from the substrate. On the opposite site is a water molecule, which was not observed in the apo GatCAB crystal, anchored by hydrogen bonds with Ser$^{153}$, Ser$^{178}$, and the amide group of the substrate itself, which is in the vicinity of hydrolysis of the enzyme-acyl intermediate. These observations clearly indicated that GatA employs the same hydrolysis mechanism as the amide hydrolysis activity of amidase. Curiously, in contrast to the previous study on GatDE from *Pyrococcus abyssi* that showed glutaminase activity only in the presence of Glu-tRNA$^{Gln}$ (*10*, *11*), GatCAB intrinsically possesses basal glutaminase activity without binding of substrate tRNA [in the current structure and by kinetics studies (*12*)], which suggests that the conformational rearrangement of the glutaminase active site of GatA is not involved in regulation of the concerted reactions.

GatB is organized into two domains connected by a characteristic ~60 Å elongated linker loop (fig. S4). The globular "cradle" domain is topologically unique and can be distinguished from GatE by two insertions on each side of the cradle (*10*). GatE has an insertion domain—resembling those found in bacterial AspRS—between Bα3 and Bβ20, and an antiparallel β sheet inserted between Bβ1 and Bα1 (fig. S3). The auxiliary C-terminal helical domain is built by an α-helical bundle architecture with seven traceable helices. The disordered C-terminal region of GatB (residues 412 to 475) is presumed to form another three-helix bundle, as indicated by the sequence identity to the C terminus of uncharacterized Yqey protein (PDB

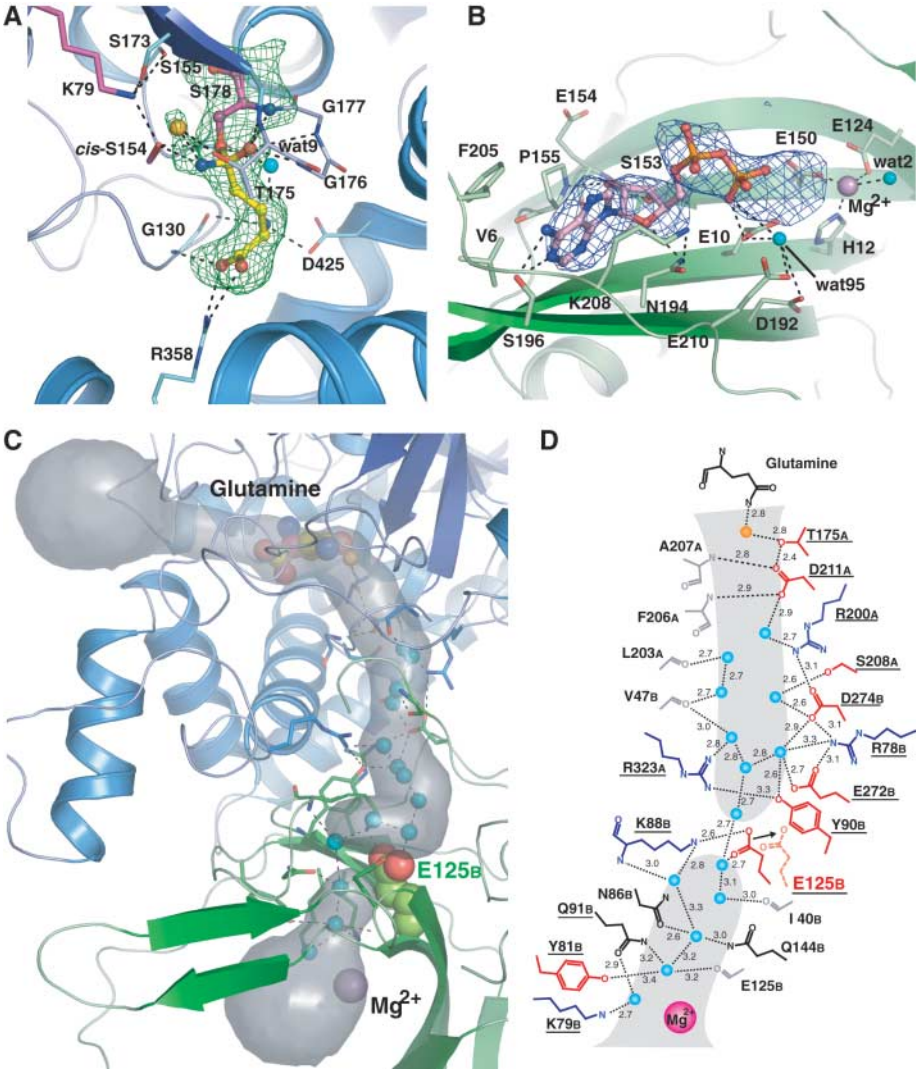



**Fig. 2.** A 30 Å long ammonia channel connects the two remote active centers of GatCAB. (**A**) The active site of a glutaminase reaction in GatA is composed of the conserved Ser-*cis*-Ser-Lys catalytic scissors shown as magenta stick representations. Residues involved in the hydrogen-bonded network (dashed black lines) in the active site are labeled. A plausible hydrolytic water molecule is colored light blue and is on the opposite side of a supposed ammonia product (orange sphere). The *Fo-Fc* electron density map (contoured at 3σ, green mesh) calculated without the glutamine and Ser$^{178}$ clearly demonstrates the tetrahedral covalent intermediate of the glutamine with Ser$^{178}$. (**B**) The environment of the ADP binding site shown together with the omit *Fo-Fc* electron density map (2σ, blue). Residues contributing to ADP (ball-and-stick) recognition are represented as stick models with labels. Two water molecules (light blue spheres) are coordinated to a magnesium ion (purple) and to β phosphate. (**C**) The putative ammonia channel was calculated using the program CAVER (*26*), with the structure of the water-omitted GatCAB/glutamine complex. Glu$^{125B}$ blocking the ammonia transport route is shown in a space-filling representation for clarity. The channel was filled with a row of solvent molecules (light blue spheres), which interact with the conserved polar residues (colored sticks) along the pathway. A bound glutamine in GatA is drawn as spheres indicating the start point of the channel. (**D**) Schematic representation of the ammonia channel. Residues defining the channel are colored corresponding to their properties: red, negative; blue, positive; black, nonpolar side chain; gray, main chain. Hydrolyzed ammonia is colored orange. Strictly conserved residues are underlined and hydrogen-bonding distances are indicated (Å). The presumed movement of the Glu125B gate to open the ammonia channel is indicated by a black arrow.

code, 1NG6). An obvious density peak of $Mg^{2+}$ ion was observed at the bottom of the cradle domain coordinated to three conserved residues consisting of $His^{12}$, $Glu^{124}$, and $Glu^{150}$, and to three water molecules (Fig. 2B). To underline the molecular basis for the mechanism of the kinase activity, we cocrystallized GatCAB with the ADP–$AlF_4^-$ compound as an ATP analog. The 3.35 Å 2*Fo-Fc* and *Fo-Fc* maps clearly indicated that ADP sits on strands Bβ1 and Bβ11 at the bottom of the cradle domain, with a bound water molecule (Fig. 2B). The adenosine base dips into a hydrophobic pocket, created by $Val^6$, $Phe^{205}$, and $Pro^{155}$. N1 and N6 are strictly recognized by an invariant $Ser^{196}$ residue in agreement with the previous report that GTP could not serve as a phosphate donor for GatCAB (*17*). The hydroxyl groups of ribose interact with the backbone of a loop between Bα2 and Bβ10, whereas the ribose O4 is recognized by the conserved Asn194 residue. The β phosphate is at an appropriate distance for hydrogen bonding with the conserved Glu10 and an anchored wat95 molecule.

Cautious inspection of the *Fo-Fc* difference electron density map from a crystal derivatized with $MnCl_2$ revealed a pair of Mn cations in the active site of GatB (fig. S4). The prominent $Mn^{2+}$ binding site is identical to the $Mg^{2+}$ site in the native crystal. The second transient binding site is located adjacent to the first (a distance of 6.3 Å), coordinated to the conserved acidic residue cluster formed by $Glu^{10}$, $Asp^{192}$, and $Glu^{210}$. This position corresponds to the wat95 water molecule in the ADP–$AlF_4^-$ structure and is at an appropriate distance to make interactions with β- and γ-phosphates of ATP. These results strongly suggested that GatB (and plausibly GatE) uses a two-metal-ion mechanism of catalysis, which is reminiscent of the mechanism employed by glutamine synthetases in amino acid biosynthesis (*18–21*). We propose that the primary $Mg^{2+}$-binding site should be responsible for coordinate bonding, recognition, and positioning of the γ-carboxyl group of the glutamate attached to the 3′ end of $tRNA^{Gln}$ (fig. S6). On the other hand, the transient $Mg^{2+}$ binding site is likely to interact with β and γ phosphates of ATP and to participate in phosphoryl transfer by polarizing the γ-phosphate group of ATP, promoting the nucleophilic attack to form a γ-phosphoryl-Glu-$tRNA^{Gln}$ (fig. S6).

The apo GatCAB structure revealed a distance of 30 Å that separates the catalytic sites at which the glutaminase reaction takes place and at which the CCA end of tRNA binds. We found that two reaction centers in GatA and GatB are connected by a penetrable tunnel lined with a continuous succession of conserved polar residues (Fig. 2, C and D). In our GatCAB/glutamine complex structure, this tunnel is thoroughly hydrophilic and is filled with a row of solvent molecules indicating the actual route of ammonia traveling from the amino acid pocket in GatA to reach the γ-phosphoryl-Glu-$tRNA^{Gln}$ in GatB. Strictly conserved positive and negative residues, designating the ammonia channel, are located alternately along the course (Fig. 2, C and D). This may imply a "proton-relay mechanism" by which ammonia is dispersed from one site to the other by repeating protonation and deprotonation with these successive conserved polar residues. This hypothesis is reconciled by the positioning of $Thr^{175A}$ at the entrance of the channel and $Lys^{79B}$ at the exit as the proton donor and acceptor, respectively. However, additional biochemical and structural studies should be required to clarify the precise mechanism of ammonia transport in GatCAB. In addition, distinct ammonia tunnels have been described in multistep enzymatic proteins that hydrolyze glutamine to obtain ammonia as a substrate for a second reaction, such as asparagine synthetase B, imidazole glycerol phosphate synthase, or glutamine phosphoribosylpyrophosphate, as well as in the case of the transmembranal ammonia transporter proteins (*22–25*). All of these adopt a dissimilar architecture of hydrophobic tunnel (~20 Å in length), suggesting that ammonia ($NH_3$), but not ammonium ion ($NH_4^+$), is transported. Therefore, GatB is idiosyncratic not only in its folding but also in its mechanism of the intramolecular ammonia transportation.

According to the program CAVER (*26*) and an in-house program used in calculation of the tunnel of GatCAB, we readily noticed a bottleneck in the middle of the path of ammonia travel (Fig. 2C). The channel is constricted by a completely conserved $Glu^{125}$ residue residing in Bβ7 of GatB, which makes a salt bridge with the invariant $Lys^{88}$ on a loop between Bβ4 and Bβ3. As the efficiency of glutamine hydrolysis is known to be stringently dependent on binding of Glu-$tRNA^{Gln}$, but neither uncharged tRNA nor Gln-$tRNA^{Gln}$ (*12*), the Glu-$tRNA^{Gln}$ recognition must induce as yet uncharacterized conformational changes. We postulate that there is an open-close motion of $Glu^{125}$ as a gate susceptible to tRNA binding. Juxtaposed to $Glu^{125}$ is

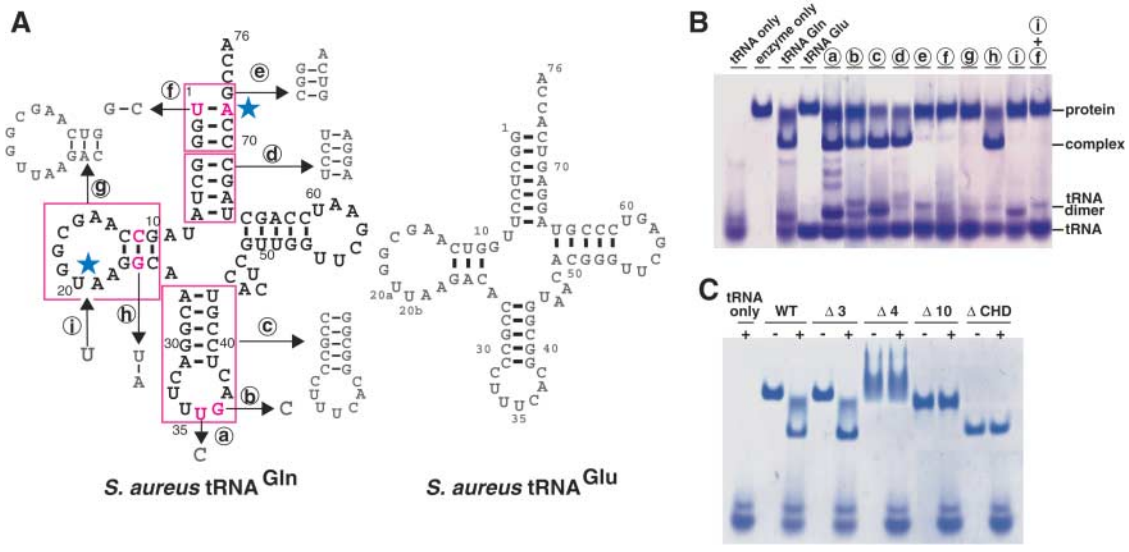



**Fig. 3.** Mutational analyses of $tRNA^{Gln}$ and GatCAB. (**A**) Cloverleaf representations of $tRNA^{Gln}$ (black) and $tRNA^{Glu}$ (gray) from *S. aureus*. Indicated bases or stem loops (magenta) were altered to those of $tRNA^{Glu}$ to verify the binding activity. Bases involved in specific recognition by GatCAB are emphasized and marked with blue stars. (**B**) Gel-shift assay with $tRNA^{Gln}$ variants as specified in (A). Bands corresponding to 150 pmol $tRNA^{Gln}$ and 50 pmol GatCAB are shown in the first and second lanes, respectively. The molecular ratios of the enzyme to $tRNA^{Gln}$ were fixed at 1:3. Uncharacterized bands above tRNA were probably responsible for the tRNA dimer suggested by the results of size-exclusion chromatography in our experiments. Gels were stained with Coomassie brilliant blue and sequentially with toluidine blue to visualize proteins and RNA, respectively. (**C**) A series of GatB deletion mutants (50 pmol GatCAB complex per lane) were subjected to gel-shift assay in the absence (–) or presence (+) of 150 pmol of the wild-type $tRNA^{Gln}$. Only $Leu^{472}$ harboring Δ3 mutant (deletion of three residues from C terminus of GatB) retained the same binding activity as wild-type (WT) GatCAB. Deletion of $Leu^{472}$ from the C terminus of GatB (Δ4 mutants) abolished the tRNA-binding capacity. Further deleted mutant (Δ10), as well as a mutant lacking the helical bundle domain (ΔCHD) of GatB, did not show any tRNA-binding activity.

a conserved $Asp^{126}$ residue, which forms a tight salt bridge with the invariant $Arg^{190}$ on the opposite Bβ11 strand, resulting in a twist in part of the Bβ7 strand. This $Asp^{126}$–$Arg^{190}$ salt bridge is located precisely at the entrance of the active site where the CCA end of tRNA will approach. It is not unreasonable to assume that the CCA end of the incoming tRNA would stimulate the movement of $Arg^{190}$, resulting in disruption of the salt bridge, relaxing the distortion of strand Bβ7, and finally opening the glutamic gate. This scenario might be a shared feature between GatCAB and GatDE, because these relevant residues are definitely conserved in both molecules.

Our structural data, however, leave the nature and specificity of binding the tRNA substrate unsolved. We constructed a series of the in vitro transcribed *S. aureus* (*Sav*) $tRNA^{Gln}$ mutants based on the differences between $tRNA^{Gln}$ and $tRNA^{Glu}$ and investigated the tRNA-binding activity with the gel-shift assay (Fig. 3, A and B). GatCAB bound quite specifically to $tRNA^{Gln}$ but not $tRNA^{Glu}$. Unexpectedly, unlike GluRS or GlnRS, GatCAB did not bind or recognize the anticodon of tRNA, whereas swapping the upper part of the acceptor stem of $tRNA^{Gln}$ to $tRNA^{Glu}$ diminished the binding activity to GatCAB. Changing only the first U1-A72 base pair of $tRNA^{Gln}$ almost extinguished the binding capacity. This is consistent with the fact that the first U1-A72 base pair is absolutely conserved and unique to $tRNA^{Gln}$ and $tRNA^{Asn}$ in bacteria (*27*) (fig. S5). No binding activity could be detected in the case of the D-arm swapped mutant. The differences in the D arms between $tRNA^{Gln}$ and $tRNA^{Glu}$ are the size of the D loop and the base pair in the middle of the D stem (Fig. 3A). Further mutational analysis of the D arm revealed that the U insertion in the D loop was lethal to tRNA recognition. These results suggested that the morphology of the D loop is important for discrimination between $tRNA^{Gln}$ and $tRNA^{Glu}$ by GatCAB. Taken together, GatCAB recognizes and discriminates two identity elements on $tRNA^{Gln}$; the U1-A72 base pair is exploited as the positive determinant to discriminate $tRNA^{Gln}$ from the other tRNA species, whereas the U insertion in the D loop serves as the negative determinant for distinguishing $tRNA^{Glu}$ from $tRNA^{Gln}$.

It is unclear how GatCAB discriminates the U1-A72 base pair in $tRNA^{Gln}$ from the G-C base pair. Interestingly, in the crystal structure of *Escherichia coli* GlnRS-$tRNA^{Gln}$ complex, the first U1-A72 base pair of $tRNA^{Gln}$ was disrupted by the $Leu^{136}$ residue at the tip of a β turn in the acceptor-binding domain of GlnRS (*28*). GatCAB may use a similar mechanism to distinguish A-U from G-C base pairs by the smaller free energy cost of base-pair denaturation. If this is the case with GatCAB, two turn loops, one between Bβ11 and Bα2 and the other between Bβ12 and Bα3, are good presumable candidates. On the other hand, we found that the C-terminal helical domain (CHD) of GatB has a crucial role in binding and discrimination of $tRNA^{Gln}$ (Fig. 3C). The mutant with deletion of the entire domain completely failed to bind to $tRNA^{Gln}$. Notably, deletion of the conserved $Leu^{472}$ residue, the fourth residue from the C terminus, of GatB also caused loss of tRNA-binding activity, whereas a GatB deletion mutant harboring $Leu^{472}$ did not impair binding activity. Because dozens of hydrophobic residues responsible for interhelical interactions are highly conserved in CHD, we concluded that elimination of $Leu^{472}$ disrupted the intact configuration of the last three helices (which were not visible in the electron density map) of CHD. This idea is corroborated by the unusually slow migration of $Leu^{472}$-deleted mutant protein in our gel-shift experiment. This distal three-helix bundle is connected to the Bα11 helix by a long, flexible linker partially seen in the present structure and expected to be able to reach and bind to the D loop of bound $tRNA^{Gln}$. Discrimination of the size or configuration of the D loop of $tRNA^{Gln}$ may be achieved by fitting to the concave made by helical bundle as a means of an "indirect readout" mechanism (*29*).

On the basis of our crystal structures combined with the evident biochemical data, we can extract a simple rationale for the operation of GatCAB controlling the Glu-$tRNA^{Gln}$–dependent amidotransferase process (fig. S6). Glu-$tRNA^{Gln}$ is successfully accommodated by GatB through two checkpoints on the body of $tRNA^{Gln}$; the U1-A72 base pair and the D loop. Subsequent precise coordination of the carboxylic acid moiety of glutamate, attached to the terminal adenosine of tRNA, to primary $Mg^{2+}$ could induce local conformational remodeling of the active site, shifting the $Glu^{125}$ gate to open the ammonia channel. This motion may slightly accelerate the glutaminase activity of GatA because of the occurrent flow of ammonia. However, the full activation depends strictly on the recruitment of ATP from the opposite site of the active center. Replacement of the wat95 molecule with the secondary $Mg^{2+}$ makes the active center ready for the transamidase reaction by preventing γ-phosphoryl-glutamate from hydrolysis. Consumption of the ammonia by the transamidase reaction would robustly drive the total potential activity of GatCAB by pumping ammonia downward with a concentration gradient. Our proposed model provides a detailed hypothesis to explain the previous biochemical and kinetic studies of the coupling reaction cycle of GatCAB (*11*, *12*); however, further biochemical and structural works on GatCAB/Glu-$tRNA^{Gln}$ complex may be needed to complete the scenario of the reaction scheme of adding glutamine to the genetic code.

### References and Notes


1. M. Ibba, D. Söll, *Science* **286**, 1893 (1999).
2. M. Wilcox, M. Nirenberg, *Proc. Natl. Acad. Sci. U.S.A.* **61**, 229 (1968).
3. M. Wilcox, *Eur. J. Biochem.* **11**, 405 (1969).
4. A. Schon, H. Hottinger, D. Söll, *Biochimie* **70**, 391 (1988).
5. M. Ibba, H. D. Becker, C. Stathopoulos, D. L. Tumbula, D. Söll, *Trends Biochem. Sci.* **25**, 311 (2000).
6. A. W. Curnow, M. Ibba, D. Söll, *Nature* **382**, 589 (1996).
7. H. D. Becker *et al.*, *FEBS Lett.* **476**, 140 (2000).
8. A. W. Curnow *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 11819 (1997).
9. D. L. Tumbula, H. D. Becker, W. Z. Chang, D. Söll, *Nature* **407**, 106 (2000).
10. E. Schmitt, M. Panvert, S. Blanquet, Y. Mechulam, *Structure* **13**, 1421 (2005).
11. L. Feng, K. Sheppard, D. Tumbula-Hansen, D. Söll, *J. Biol. Chem.* **280**, 8150 (2005).
12. K. Y. Horiuchi *et al.*, *Biochemistry* **40**, 6450 (2001).
13. M. Yao, Y. Zhou, I. Tanaka, *Acta Crystallogr. D Biol. Crystallogr.* **62**, 189 (2006).
14. A. T. Brunger *et al.*, *Acta Crystallogr. D* **54**, 905 (1998).
15. Materials and methods are available as supporting material on *Science* Online.
16. S. Shin *et al.*, *EMBO J.* **21**, 2509 (2002).
17. D. Jahn, Y. C. Kim, Y. Ishino, M. W. Chen, D. Söll, *J. Biol. Chem.* **265**, 8059 (1990).
18. F. C. Wedler, P. D. Boyer, *J. Biol. Chem.* **247**, 984 (1972).
19. J. B. Hunt, A. Ginsburg, *Biochemistry* **11**, 3723 (1972).
20. E. R. Stadtman *et al.*, *Mol. Biol. Biochem. Biophys.* **32**, 144 (1980).
21. D. Eisenberg, H. S. Gill, G. M. Pfluegl, S. H. Rotstein, *Biochim. Biophys. Acta* **1477**, 122 (2000).
22. T. M. Larsen *et al.*, *Biochemistry* **38**, 16146 (1999).
23. A. Douangamath *et al.*, *Structure* **10**, 185 (2002).
24. A. K. Bera, J. L. Smith, H. Zalkin, *J. Biol. Chem.* **275**, 7975 (2000).
25. S. Khademi *et al.*, *Science* **305**, 1587 (2004).
26. O. M. Petrek, P. Banas, J. Koca, J. Damborsky, CAVER program, http://loschmidt.chemi.muni.cz/caver/index.php.
27. C. Marck, H. Grosjean, *RNA* **8**, 1189 (2002).
28. M. A. Rould, J. J. Perona, D. Söll, T. A. Steitz, *Science* **246**, 1135 (1989).
29. S. Hauenstein, C. M. Zhang, Y. M. Hou, J. J. Perona, *Nat. Struct. Mol. Biol.* **11**, 1134 (2004).


30. We thank Y. Tanaka for helpful advice in initial molecular cloning and purification processes of GatCAB complex and the staff of beamlines BL41XU and BL44B2 at SPring-8, as well as of ARNW12 station at Proton Factory, for help during data collection. We are grateful to O. Nureki (Tokyo Institute of Technology) for providing the refined coordinates of GatDE/$tRNA^{Gln}$ complex before publication. This work was supported by the National Project on Protein Structural and Functional Analyses from the Ministry of Education, Culture, Sports, Sciences, and Technology of Japan and was supported in part by a Human Frontiers Science Program Research grant. Coordinates and structure-factor amplitudes have been deposited in the Protein Data Bank under codes 2G5H, 2F2A, 2G5I, 2DF4, and 2DQN for the apo form and the glutamine-bound, ATP analog-bound, $Mn^{2+}$-bound, and asparagine-bound forms, respectively.


### Supporting Online Material

www.sciencemag.org/cgi/content/full/312/5782/1954/DC1
Materials and Methods
Figs. S1 to S6
Table S1
References

8 March 2006; accepted 30 May 2006
10.1126/science.1127156

# Rapid Advance of Spring Arrival Dates in Long-Distance Migratory Birds

Niclas Jonzén,[1] Andreas Lindén,[2] Torbjørn Ergon,[4] Endre Knudsen,[4] Jon Olav Vik,[4] Diego Rubolini,[5] Dario Piacentini,[6] Christian Brinch,[4] Fernando Spina,[6] Lennart Karlsson,[7] Martin Stervander,[8] Arne Andersson,[8] Jonas Waldenström,[9] Aleksi Lehikoinen,[3] Erik Edvardsen,[10] Rune Solvang,[10] Nils Chr. Stenseth[4]*

Several bird species have advanced the timing of their spring migration in response to recent climate change. European short-distance migrants, wintering in temperate areas, have been assumed to be more affected by change in the European climate than long-distance migrants wintering in the tropics. However, we show that long-distance migrants have advanced their spring arrival in Scandinavia more than short-distance migrants. By analyzing a long-term data set from southern Italy, we show that long-distance migrants also pass through the Mediterranean region earlier. We argue that this may reflect a climate-driven evolutionary change in the timing of spring migration.

Many biological processes are affected by climate, and in temperate areas the increasing spring temperature over the past 20 to 30 years has caused an advancement of phenological events in plants and invertebrates (*1*, *2*). The earlier onset of spring has consequences for the timing of breeding in birds, which has evolved to match peak food availability (*3*, *4*). We may therefore expect the timing of breeding to track any temporal shift in food availability caused by a trend in spring temperature (*5*). Most passerine birds breeding in temperate areas of the Northern Hemisphere are seasonal migrants, and the timing of migration ultimately constrains when breeding can start (*6*, *7*). Short-distance migrants, spending the winter close to the breeding grounds, may be able to adjust the timing of migration in response to local climate change, which will be correlated to the conditions on the breeding grounds. In tropical-wintering long-distance migrants, the timing of migration is under endogenous control (*8*, *9*), and the cues needed to trigger the onset of migration are unlikely to be linked to the climate on their breeding grounds. Therefore, it has been assumed that short-distance migrants are more likely than long-distance migrants to vary migration timing in response to climate change (*10*). Here we show that such an assumption is not empirically justified.

We estimated trends in arrival time for the early, middle, and late phases of migration (that is, the species- and site-specific 10th, 50th, and 90th percentiles of the spring arrival distribution) in short- and long-distance passerine migrants, based on long-term banding and observational data (from 1980 to 2004) from four bird observatories in Scandinavia and a site in southern Italy (*11*, Fig. 1). We also investigated whether year-to-year variation in arrival time can be explained by short-term climate variability as measured by the North Atlantic Oscillation (NAO) (*12*). As explanatory variables we used the calendar year (TIME) and the deviations from linear regression of the winter NAO index on year [dNAO; the trend in NAO was weakly negative over this time period (*11*)]. Spring migration might advance for two distinct reasons. First, there can be a microevolutionary (genetic) response to the selection pressures for earlier breeding. Second, the migrants can show a phenotypically plastic response to trends in weather or climatic patterns on the wintering ground and/or along the migration route, whereby if spring arrives early on the wintering grounds, spring migration will also start early. Thus, a response to TIME may reflect either microevolutionary change or phenotypic plasticity, whereas a response to dNAO indicates exclusively phenotypic plasticity in the migratory behavior.

Long-distance migrants have advanced their arrival in northern Europe in all phases of migration (Fig. 2 and tables S1 to S3). The advancement in long-distance migrants is strongest in the early phase of migration, and there is limited variation between species. Furthermore, the analysis of the data set from Italy (from the island of Capri) showed that long-distance migrants wintering south of the Sahara desert are actually arriving in southern Europe progressively earlier. In fact, all of the nine species analyzed show a trend for earlier spring arrival at Capri in most phases of migration (Fig. 2 and table S4). The long-term trend on Capri is at least as strong as that observed in Scandinavia (Fig. 3). In short-distance migrants, instead, we find only a weak trend toward earlier arrival, and there is considerable variation between species (Fig. 2 and tables S1 to S3).

In accordance with previous findings (*13*–*15*), a high NAO index is associated with the early arrival of short-distance migrants in Scandinavia, but only in the early phase of migration (Fig. 2).

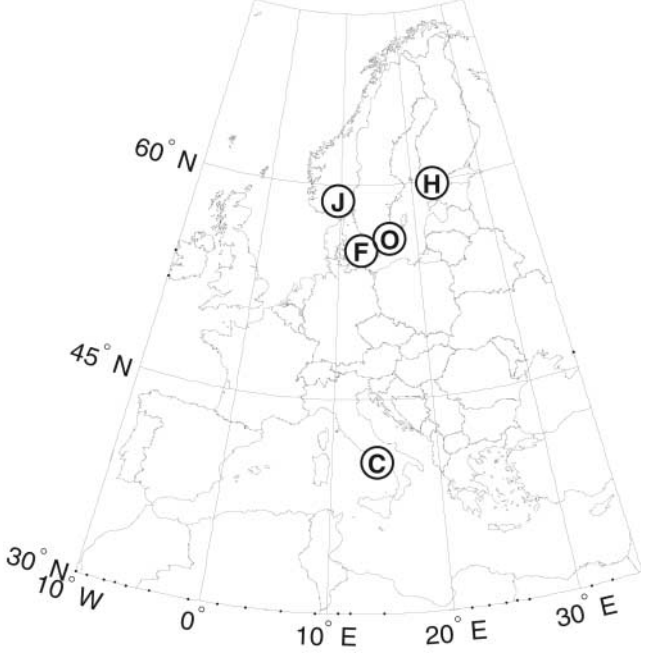



**Fig. 1.** The locations of the four bird observatories (F, Falsterbo, 55°23′N, 12°49′E; O, Ottenby, 56°12′N, 16°24′E; J, Jomfruland, 58°53′N, 9°37′E; H, Hanko, 59°48′N, 22°53′E) and of the banding site on Capri (C, 40°33′N, 14°15′E).

[1]Department of Theoretical Ecology, Ecology Building, Lund University, SE-22362 Lund, Sweden. [2]Department of Biological and Environmental Sciences, Integrative Ecology Unit; [3]Department of Biological and Environmental Sciences, Bird Ecology Unit; Post Office Box 65 (Viikinkaari 1), FIN-00014, Helsinki University, Finland. [4]Centre for Ecological and Evolutionary Synthesis, Department of Biology, University of Oslo, Post Office Box 1066, Blindern, N-0316 Oslo, Norway. [5]Dipartimento di Biologia Animale, Università degli Studi di Pavia, Piazza Botta 9, I-27100 Pavia, Italy. [6]Istituto Nazionale per la Fauna Selvatica, via Ca′ Fornacetta 9, I-40064 Ozzano Emilia (BO), Italy. [7]Falsterbo Bird Observatory, Fyren, SE-23940 Falsterbo, Sweden. [8]Ottenby Bird Observatory, Post Office Box 1500, SE-38065 Degerhamn, Sweden. [9]Section for Zoonotic Ecology and Epidemiology, Department of Biology and Environmental Sciences, Kalmar University, SE-39182 Kalmar, Sweden. [10]Jomfruland Bird Observatory, Post Office Box 1076, N-3704 Skien, Norway.

*To whom correspondence should be addressed. E-mail: n.c.stenseth@bio.uio.no

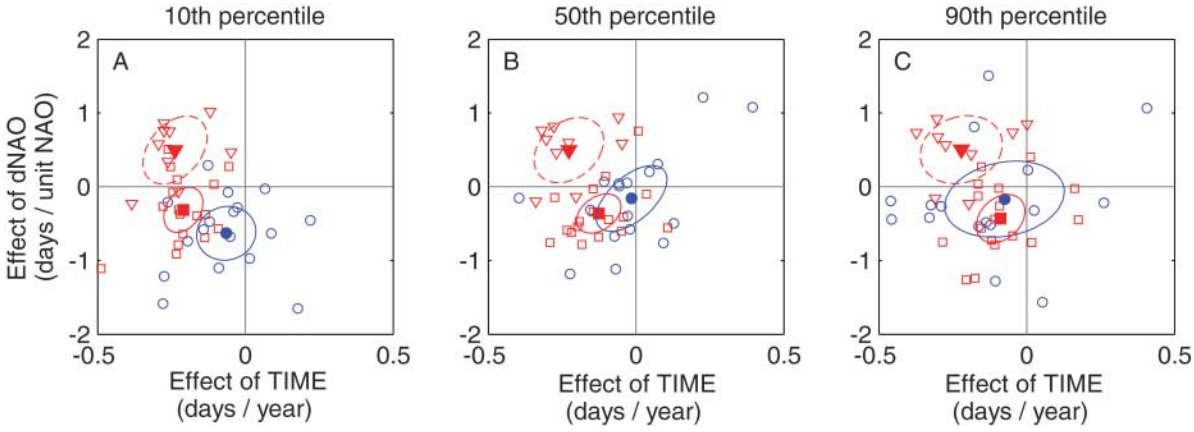



**Fig. 2.** Long-term trend (TIME) and the effect of short-term climatic fluctuations (dNAO) on the early [(**A**), 10th percentile], middle [(**B**), 50th percentile], and late [(**C**), 90th percentile] phases of the spring arrival distribution in short-distance migrants (blue circles) and long-distance migrants (red squares) in Scandinavia and on Capri (red triangles). The solid symbols are sample averages, and the ellipses delimit their 95% confidence regions (*11*). Estimates for each species are given in tables S1 to S4. The differences in effect size for early-phase arrival of short-distance migrants versus long-distance migrants was 0.15 [95% confidence interval (CI): 0.06 to 0.23] days per year for the effect of TIME.

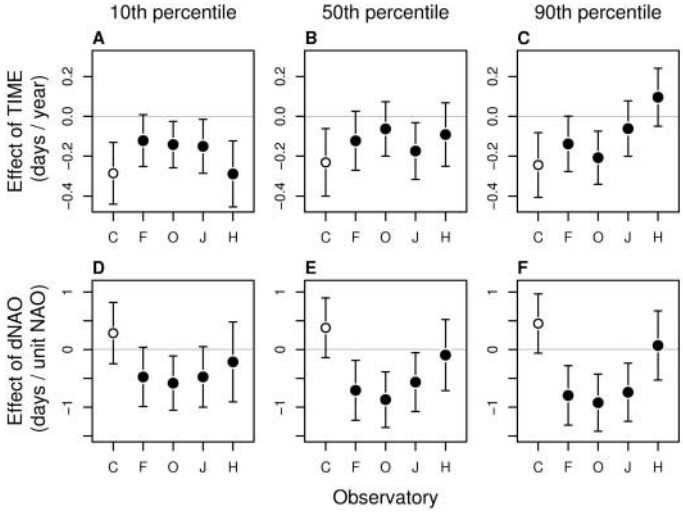



**Fig. 3.** Mean slopes and 95% CIs of the relationships between arrival dates and TIME (**A** to **C**) or dNAO (**D** to **F**) in the early (10th percentile), middle (50th percentile), and late (90th percentile) phases of migration for six species of long-distance migrants for which time series were available at each observatory (solid symbols; F, Falsterbo; O, Ottenby; J, Jomfruland; H, Hanko) and at the banding site on Capri (open symbol; C). The sites are sorted from south to north. Species-specific slopes were estimated with a mixed-effect linear model (*11*). The correlation between species, when species- and observatory-specific effects of both TIME and dNAO were accounted for, was estimated at 0.48 (95% CI: 0.17 to 0.75), 0.25 (95% CI: 0.04 to 0.51), and 0.34 (95% CI: 0.03 to 0.68) for the 10th, 50th, and 90th percentiles, respectively.

On the other hand, most long-distance migrants tend to arrive earlier in Scandinavia during years of high NAO in all phases of migration (Figs. 2 and 3 and tables S1 to S3). The opposite pattern is observed at Capri, where high NAO tends to delay arrival times (Figs. 2 and 3 and data in table S4). The underlying reason for this may be found south of the Sahara desert, because a high NAO index harms productivity over vast areas of northwestern and southeastern Africa (*16*), which may delay the spring departure of migrants from sub-Saharan wintering areas.

By showing that long-distance migrants have advanced their migration more than short-distance migrants, we have challenged the conventional wisdom that species wintering in temperate Europe should respond more strongly to climate change than trans-Saharan migrants (*10*). Furthermore, the earlier arrival of trans-Saharan migrants at Capri shows that the temporal trend for earlier arrival in Scandinavia cannot be explained simply by faster migration through Europe in response to a concomitant trend of increasing temperatures taking place within continental Europe (*17*). Instead it suggests that (i) the onset of migration has advanced, or (ii) the speed of migration through Africa has increased. Both alternatives could be seen as phenotypic responses to trends in the African climate patterns having a positive effect on the foraging conditions (*18*), thereby improving the birds' physical conditions, which in turn affects their timing of migration (*19*) and makes the migration (including flight and stopover) more efficient. A positive trend in African temperatures (*20*) has previously been suggested as a reason why long-distance migrants arrive earlier in northern Europe (*21*). However, increasing African temperatures should decrease productivity (*22*), thereby delaying long-distance migrants' departure from the wintering ground. Hence, the earlier arrival is probably not a phenotypic response to improved foraging conditions. More likely, the rapid advance in arrival dates of long-distance migrants in Europe is due to climate-driven evolutionary changes in the timing of spring migration. Even though migratory activity is under endogenous control, experiments have demonstrated individual variation in the response to the photoperiodic cues needed to trigger the mechanisms underlying the onset of migration (*23*). The passerine birds investigated here reproduce at just 1 year of age and thus have the potential for a rapid evolutionary response to environmental changes. Given the considerable heritable genetic variation in the timing of migration (*24*, *25*) and the selection pressure to breed earlier in Europe (*6*, *7*), a change toward earlier arrival is indeed to be expected.

### References and Notes

1. R. Harrington, I. Woiwood, T. H. Sparks, *Trends Ecol. Evol.* **14**, 146 (1999).
2. T. L. Root *et al.*, *Nature* **421**, 57 (2003).
3. D. Lack, *Ecological Adaptations for Breeding in Birds* (Methuen, London, 1968).
4. M. E. Visser, C. Both, M. M. Lambrecht, *Adv. Ecol. Res.* **35**, 89 (2004).
5. P. O. Dunn, *Adv. Ecol. Res.* **35**, 69 (2004).
6. C. Both, M. E. Visser, *Nature* **411**, 296 (2001).
7. C. Both, S. Bouwhuis, C. M. Lessels, M. E. Visser, *Nature* **441**, 81 (2006).
8. P. Berthold, *Control of Bird Migration* (Cambridge Univ. Press, Cambridge, 1996).
9. E. Gwinner, *J. Exp. Biol.* **199**, 39 (1996).
10. E. Lehikoinen, T. H. Sparks, M. Zalakevicius, *Adv. Ecol. Res.* **35**, 1 (2004).
11. The details of the data and methods are available as supporting material on *Science* Online.

12. J. W. Hurrell, *Science* **269**, 676 (1995).
13. O. Hüppop, K. Hüppop, *Proc. R. Soc. London Ser. B* **270**, 233 (2003).
14. M. Stervander, Å. Lindström, N. Jonzén, A. Andersson, *J. Avian Biol.* **36**, 210 (2005).
15. A. V. Vähätalo, K. Rainio, A. Lehikoinen, E. Lehikoinen, *J. Avian Biol.* **35**, 210 (2004).
16. L. C. Stige *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 3049 (2006).
17. C. Both, R. G. Bijlsma, M. E. Visser, *J. Avian Biol.* **36**, 368 (2005).
18. N. Saino *et al.*, *Ecol. Lett.* **7**, 21 (2004).
19. P. P. Marra, K. A. Hobson, R. T. Holmes, *Science* **282**, 1884 (1998).
20. M. Hulme, R. Doherty, T. Ngara, M. New, D. Lister, *Clim. Res.* **17**, 145 (2001).
21. P. A. Cotton, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 12219 (2003).
22. O. Gordo, L. Brotons, X. Ferrer, P. Comas, *Global Change Biol.* **11**, 12 (2005).
23. T. Coppack, F. Pulido, M. Czisch, D. P. Auer, P. Berthold, *Proc. R. Soc. London Ser. B* **270**, S43 (2003).
24. A. P. Møller, *Proc. R. Soc. London Ser. B* **268**, 203 (2001).
25. F. Pulido, P. Berthold, in *Avian Migration*, P. Berthold, E. Gwinner, E. Sonnenschein, Eds. (Springer-Verlag, Berlin, 2003), pp. 53–77.
26. Funding for the analysis reported in this paper was provided by the Nordic Council through the NCoE-EcoClim and the Swedish Research Council (to N.J.). Funding for obtaining the Ottenby data was provided by the Swedish Environmental Protection Agency. This is contribution number 215 from Ottenby Bird Observatory, contribution number 232 from Falsterbo Bird Observatory, contribution number 79 from Jomfruland Bird Observatory, and results from the Progetto Piccole Isole (Istituto Nazionale per la Fauna Selvatica), paper no. 37.

**Supporting Online Material**
www.sciencemag.org/cgi/content/full/312/5782/1959/DC1
Methods
SOM Text
Tables S1 to S5
References

13 February 2006; accepted 25 May 2006
10.1126/science.1126119

# Intron Removal Requires Proofreading of U2AF/3′ Splice Site Recognition by DEK

**Luis Miguel Mendes Soares,[1] Katia Zanier,[4] Cameron Mackereth,[4] Michael Sattler,[4] Juan Valcárcel[1,2,3]***

**Discrimination between splice sites and similar, nonsplice sequences is essential for correct intron removal and messenger RNA formation in eukaryotes. The 65- and 35-kD subunits of the splicing factor U2AF, U2AF$^{65}$ and U2AF$^{35}$, recognize, respectively, the pyrimidine-rich tract and the conserved terminal AG present at metazoan 3′ splice sites. We report that DEK, a chromatin- and RNA-associated protein mutated or overexpressed in certain cancers, enforces 3′ splice site discrimination by U2AF. DEK phosphorylated at serines 19 and 32 associates with U2AF$^{35}$, facilitates the U2AF$^{35}$-AG interaction and prevents binding of U2AF$^{65}$ to pyrimidine tracts not followed by AG. DEK and its phosphorylation are required for intron removal, but not for splicing complex assembly, which indicates that proofreading of early 3′ splice site recognition influences catalytic activation of the spliceosome.**

A minimal U2AF heterodimer consisting of RNA recognition motifs (RRM) 1 and 2 of U2AF$^{65}$ (*1*) and the U2AF homology motif (UHM or ΨRRM) of U2AF$^{35}$ (*2*) was analyzed by nuclear magnetic resonance (NMR) spectroscopy in the absence or presence of an RNA containing a pyrimidine tract followed by a consensus 3′ splice site (3′ss) [5′ $(U)_{13}$ACAGG 3′]. As expected from the affinity of U2AF$^{65}$ for uridine-rich sequences (*1*), the presence of the RNA caused extensive changes in the NMR spectrum of the U2AF$^{65}$ RRM 1+2 subunit (Fig. 1A, left). In contrast, small perturbations concerning few residues were observed in the U2AF$^{35}$ ΨRRM spectrum (right). The latter was unexpected, because previous observations suggested that U2AF$^{35}$ specifically recognizes the 3′ss-AG (*3*–*5*). Gel retardation assays using $^{32}$P-uridine–labeled RNAs [5′ GGG$(U)_{13}$AC-AG/CG-GUAAAAUAACUCA 3′] showed that, although U2AF$^{35}$ ΨRRM increases the affinity of the complex threefold, the effect is similar for AG-, CG-, UG- or AA-3′ss, strong or weaker pyrimidine tracts (Fig. 1B and figs. S1 and S2). Lack of AG discrimination was also observed when different assays and recombinant full-length U2AF heterodimer or U2AF purified from HeLa cells were utilized (Figs. 1, C and D). In contrast, both endogenous U2AF and the minimal heterodimer showed preferential ultraviolet (UV) light–induced photo–cross-linking of U2AF$^{65}$ to AG-3′ss RNAs in nuclear extracts (Fig. 1E). Reconstitution of U2AF-depleted extracts with recombinant U2AF subunits indicated that U2AF$^{35}$ is required for AG discrimination (Fig. 1E, bottom). The presence of U2AF$^{35}$ and other components of the nuclear extract decreased cross-linking of U2AF$^{65}$ to the nonconsensus CG-3′ss, which suggests the existence of a proofreading activity that enforces specific association of U2AF with pyrimidine tracts followed by consensus AG-3′ss.

This activity cofractionated with U2AF during the two first chromatographic steps of U2AF purification (*6*) (fig. S3). In fig. 2A, compare lanes 3 and 4 with 7 and 8 for the U2AF-containing complex (identified in lane 2 by supershift with antibodies against U2AF$^{65}$). The activity was, however, separated from U2AF on the next chromatographic step [poly(U)-Sepharose]; whereas U2AF was retained in the column (*6*), the flow-through fraction provided AG versus CG discrimination to the truncated heterodimer in both UV-mediated cross-linking (Fig. 2B) and gel-retardation assays (fig. S4). The activity present in this fraction was retained on an affinity column containing the truncated U2AF heterodimer (Fig. 2D, lanes 1 to 4). Comparison of the protein profiles of the input and flow-through fractions revealed that a 50-kD protein was retained in the U2AF column (Fig. 2C, lower component of the 50-kD doublet). Mass spectrometry analyses identified this protein as DEK, a chromatin-, pre–mRNA- and mRNA-associated protein overexpressed or mutated in certain cancers (*7*, *8*). Consistent with a role for DEK in providing AG discrimination to U2AF, depletion of DEK from HeLa nuclear extracts (fig. S5) resulted in reduced AG versus CG discrimination by endogenous U2AF$^{65}$ (Fig. 2E, lanes 1 to 4), an effect that was reversed when recombinant purified DEK was added to the depleted extracts (lane 5). Cross-linking between U2AF$^{35}$ and an RNA radioactively labeled at the 3′ss dinucleotide (A-[$^{32}$P]-G or C-[$^{32}$P]-G) was reduced in DEK-depleted extracts, which indicated that DEK is required for 3′ss recognition by U2AF$^{35}$ (Fig. 2F). Collectively, the results described above indicate that DEK provides a proofreading function that allows U2AF to discriminate between bona fide AG-containing and nonconsensus 3′ss regions.

DEK retention in U2AF affinity columns suggested the possibility of an interaction between these factors. Pull-down experiments using in vitro translated, $^{35}$S-labeled U2AF$^{65}$ or U2AF$^{35}$ and recombinant purified glutathione *S*-transferase (GST)-DEK revealed formation of a complex between DEK and U2AF$^{35}$, which was, at least in part, RNA-independent and involved the 100 amino-terminal residues of DEK (Fig. 3, A and B). Interestingly, the interaction was disrupted by phosphatase treatment (Fig. 3B, lanes 3 versus 4 and 11 versus 12), which suggests the requirement for protein phosphorylation. Indeed, DEK is a phospho-

[1]Centre de Regulació Genòmica, [2]Institució Catalana de Recerca i Estudis Avançats, [3]Universitat Pompeu Fabra, Passeig Marítim 37-49, 08003 Barcelona, Spain. [4]European Molecular Biology Laboratory, Meyerhofstrasse 1, D-69117 Heidelberg, Germany.

*To whom correspondence should be addressed. E-mail: juan.valcarcel@crg.es

protein (*9*), and phosphorylation occurs upon incubation of DEK(1–100) with rabbit reticulocyte lysates (fig. S7). Mutation of serines 19 and 32 to alanine (A) or aspartic acid (D) abolished phosphorylation (fig. S6). Interaction with U2AF$^{35}$ was abolished by mutation to alanine; mutation to the phosphorylation mimic, aspartic acid, allowed the interaction to occur even in the presence of phosphatase (Fig. 3B, lanes 5 to 8 and 13 to 16). Taken together, the results of Fig. 3 reveal formation of a complex involving the amino-terminal 100 amino acids of DEK and U2AF$^{35}$, which is dependent on phosphorylation of serines 19 and 32. The involvement of additional factors in complex formation cannot be ruled out, because NMR experiments failed to detect significant DEK-induced changes in U2AF$^{35}$ structure or interaction with RNA.

Recombinant DEK was found insufficient to confer discrimination to the purified heterodimer (Fig. 2D, lanes 5 and 6). Discrimination was restored, however, when DEK was combined with the DEK-depleted flow-through fraction from the U2AF affinity column mentioned above (Fig. 2D, lanes 7 and 8), but not with other chromatographic fractions (fig. S3B). The requirement for adenosine triphosphate (ATP) (see below), for U2AF$^{35}$ (Fig. 2D, lanes 9 and 10) and the phosphorylation-dependent association of U2AF$^{35}$ suggested the possibility that DEK phosphorylation by a kinase present in the complementing fraction was necessary for AG discrimination by U2AF. Consistent with this hypothesis, DEK was phosphorylated in the presence of the complementing fraction (fig. S6), and mutation of serines 19 and 32 to alanine compromised the discriminatory activity provided by DEK, whereas mutation to aspartic acid maintained the activity in reconstituted assays (Fig. 2D, lanes 11 to 14) and in complementation of DEK-depleted extracts (Fig. 2E, lanes 6 and 7). Part of the discrimination provided by the mutant with aspartic acid at positions 19 and 32 was lost in the absence of ATP (lanes 15 and 16), which suggests that phosphorylation of serines 19 and 32 is necessary, but not sufficient, for full DEK activity. Collectively, the results of Figs. 2 and 3 indicate that DEK phosphorylation facilitates its association with U2AF$^{35}$, and this correlates with the activity of the protein to confer AG discrimination to the U2AF heterodimer.

One possible consequence of the absence of DEK proofreading would be splicing activation of pre-mRNAs containing mutations at the 3′ss AG. This was found not to be the case for either IgM pre-mRNA (which requires the 3′ss AG to undergo both catalytic steps of splicing) or AdML (which undergoes—at least to some extent—the first catalytic step in the absence of 3′ss AG) (Fig. 4A). DEK depletion, however, inhibited the catalytic steps requiring AG recognition in wild-type RNAs: first step for the AG-dependent and, mainly, second step for the AG-independent pre-mRNAs (Fig. 4B). Splicing was restored by complementation with recombinant purified DEK and the DEK mutant with Asp$^{19}$ and Asp$^{32}$, but not with the DEK mutant with Ala$^{19}$ and Ala$^{32}$ (Fig. 4, B and C, and fig. S7). U2AF was not depleted in DEK-depleted extracts (fig. S5), and splicing was not restored by an excess of U2AF$^{65}$ or U2AF heterodimer (Fig. 4C). Collectively, these data establish a correlation—further substantiated by the common requirement for phosphorylation of Ser$^{19}$ and Ser$^{32}$—between three properties of DEK: associating with U2AF35, proofreading U2AF/3′ ss recognition, and sustaining pre-mRNA splicing.

In contrast, neither U2 small nuclear ribonucleoprotein (snRNP) binding nor subsequent events in spliceosome assembly were affected by DEK depletion (Fig. 4D), which suggests that the absence of proofreading at early stages of 3′ss recognition influences catalytic activation of splicing complexes. Effects of DEK depletion on splicing complex formation were, however, observed in competition assays in which U2 snRNP binding to a 3′ss-containing RNA was competed by an excess of RNAs containing a pyrimidine tract followed by AG or CG. Although only AG-containing RNAs

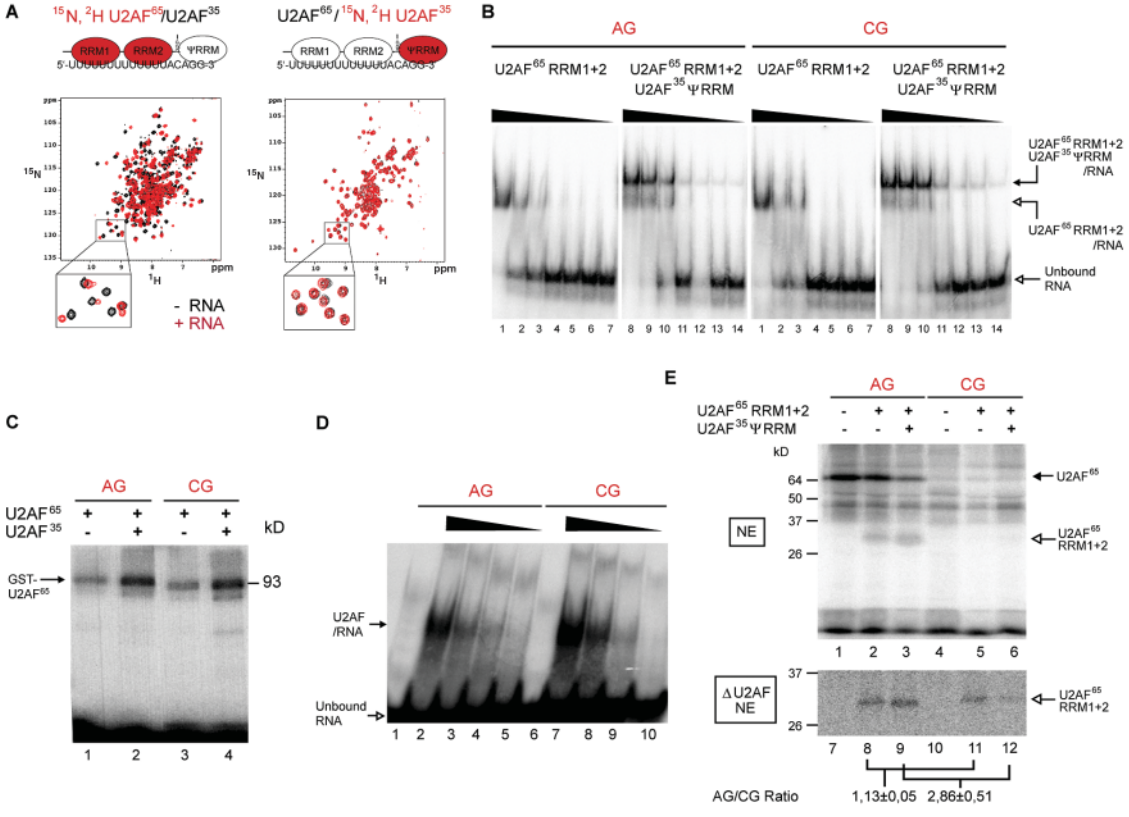



**Fig. 1.** An activity present in nuclear extracts is necessary for 3′ss AG discrimination by U2AF. (**A**) Two-dimensional NMR spectra of $^{1}$H,$^{15}$N amide resonance correlations of U2AF$^{65}$ RRM 1+2 (left) and U2AF$^{35}$ ΨRRM (right) forming a minimal U2AF heterodimer. Spectra in the absence (black) or presence (red) of a twofold molar excess of 5′(U)$_{13}$ACAGG 3′ RNA are shown. (**B**) Gel retardation assays using the indicated proteins ($10^{-6}$ M for lanes 1 and 8, and serial threefold dilutions) and 10 fmol of either 5′ $^{32}$P-labeled RNA used in (A) (AG) or a mutant replacing the 3′ss AG by CG (CG). Positions of unbound RNA and complexes are indicated. (**C**) UV-induced cross-linking assay using GST-U2AF$^{65}$ ($2 \times 10^{-7}$ M), His-U2AF$^{35}$ ($5 \times 10^{-7}$ M) and $^{32}$P-uridine–labeled 5′ GGG(U)$_{13}$AC-AG/CG-GUAAAAUAACUCA 3′ RNAs ($10^{-9}$ M). (**D**) Gel retardation assay using purified U2AF heterodimer from HeLa cells ($10^{-8}$ M for lanes 2 and 7, and serial twofold dilutions). (**E**) UV-induced cross-linking and immunoprecipitation of endogenous U2AF and truncated heterodimer subunits ($2.5 \times 10^{-5}$ M) in the presence of HeLa (NE) or U2AF-depleted (ΔU2AF NE) nuclear extracts (33% of reaction volume). Ratios between cross-linking intensities for AG- and CG-containing RNAs are indicated ($n = 3$).

were able to compete in mock-depleted extracts, both AG- and CG-containing RNAs were effective competitors in DEK-depleted extracts (Fig. 4E and fig. S8). These results are consistent with the notion that DEK enforces selective functional association of U2AF with bona fide 3′ss.

Alterations in splice-site recognition and intron removal could underlie the molecular basis of pathologies involving DEK overexpression or inactivation, including autoimmune disease and cancer (*7*). Consistent with this, overexpression of mutant versions of DEK in cells in culture results in reduced splicing efficiency of model pre-mRNAs (fig. S9). Previous data from iterative selection of sequences from a random pool (*3*) indicated that the U2AF heterodimer is sufficient to select both pyrimidine tracts and consensus 3′ss AG. DEK may enhance this intrinsic preference, revealed to different extents depending on assay stringency. DEK binding could induce a conformational change in U2AF$^{35}$ that enforces AG selectivity (Fig. 2F), similar to the enhanced specificity of U2A′ for U2 snRNA sequences upon interaction with U2B″ (*10*). Elegant studies in yeast implicated the transient interaction of the RNA-

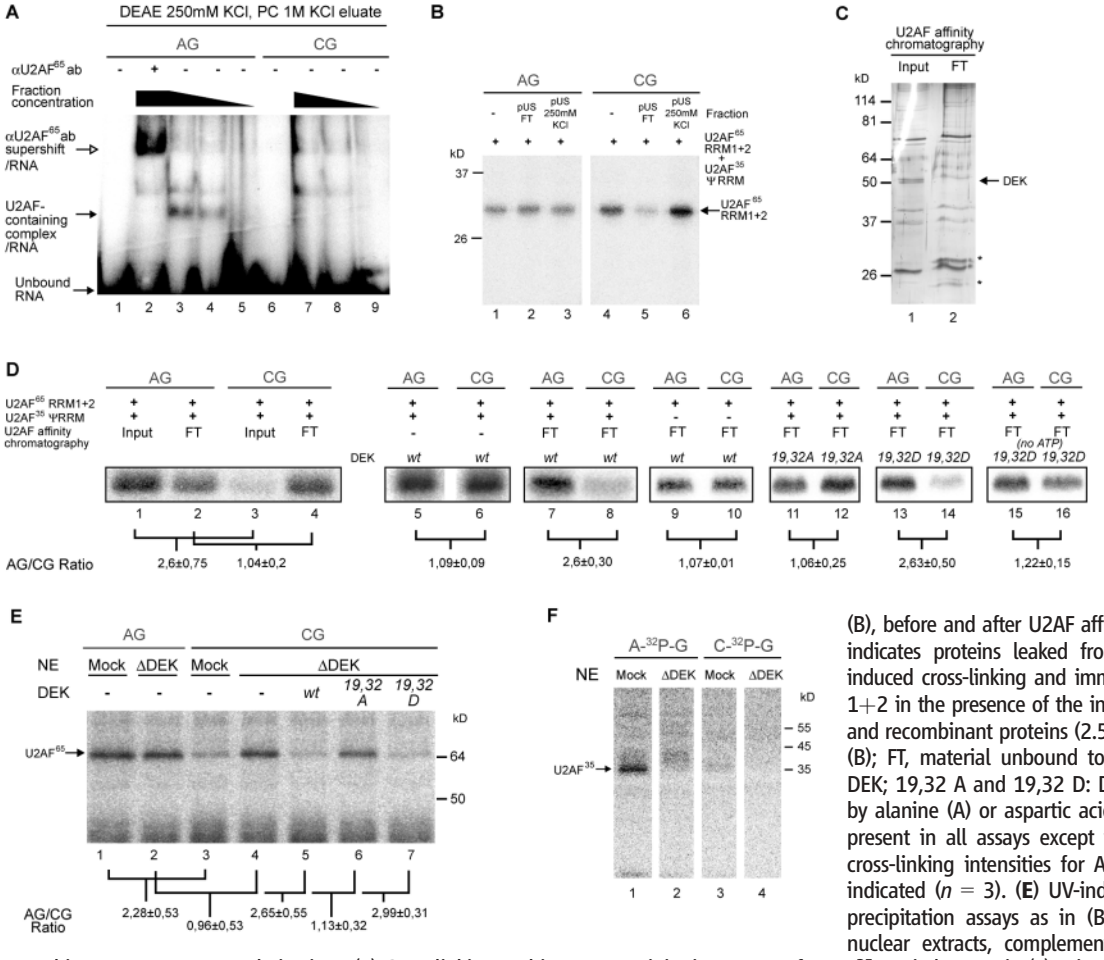



**Fig. 2.** DEK enforces discrimination by U2AF between 3′ss AG and CG. (**A**) Gel retardation assay using the DEAE Sepharose, phosphocellulose (PC) eluate of U2AF purification procedure (*6*) carried out as in Fig. 1B. The position of the U2AF-containing complex (as shown by supershift with antibody against U2AF$^{65}$) is indicated. (**B**) UV-induced cross-linking and immunoprecipitation of U2AF$^{65}$ RRM 1+2 in the presence of the indicated chromatographic fractions, carried out as in Fig. 1E. pUS FT, flow-through of poly(U) Sepharose loaded with the fraction tested in (A). pUS 250 mM KCl, eluate of the same column at 250 mM KCl. (**C**) Silver stain of protein profiles of chromatographic fractions described in (B), before and after U2AF affinity chromatography. The asterisk indicates proteins leaked from the affinity column. (**D**) UV-induced cross-linking and immunoprecipitation of U2AF$^{65}$ RRM 1+2 in the presence of the indicated chromatographic fractions and recombinant proteins ($2.5 \times 10^{-6}$ M). Input: pUS FT, as in (B); FT, material unbound to U2AF affinity column; wt, GST-DEK; 19,32 A and 19,32 D: DEK mutants with serine replaced by alanine (A) or aspartic acid (D) at these positions. ATP was present in all assays except where indicated. Ratios between cross-linking intensities for AG- and CG-containing RNAs are indicated ($n = 3$). (**E**) UV-induced cross-linking and immunoprecipitation assays as in (B), using mock- or DEK-depleted nuclear extracts, complemented with 100 ng/μl of purified recombinant DEK or mutant derivatives. (**F**) Cross-linking and immunoprecipitation assays of U2AF$^{35}$ carried out as in (E) using RNAs $^{32}$P-labeled at ApG or CpG 3′ss and antibodies against U2AF$^{35}$.

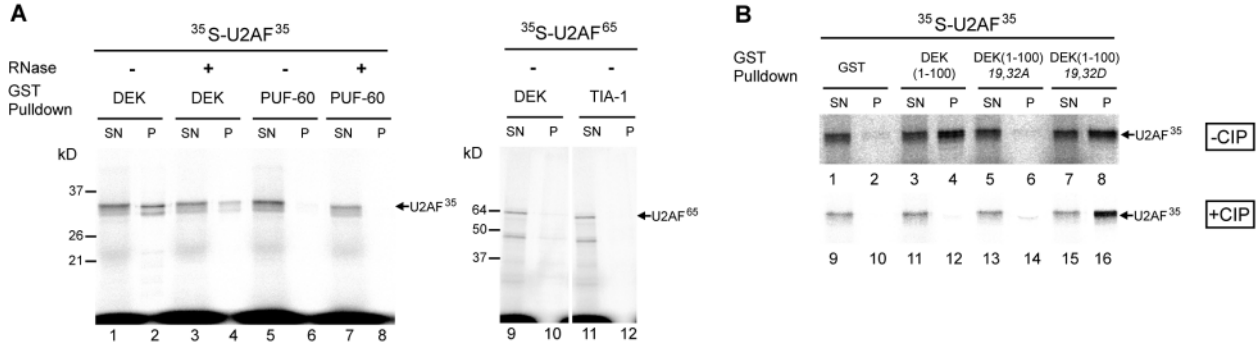



**Fig. 3.** Phosphorylation-dependent association between DEK and U2AF$^{35}$. (**A**) Pull-down of $^{35}$S-labeled in vitro translated U2AF subunits by the indicated GST fusion proteins (PUF-60 and TIA-1 are splicing factors used as negative controls). P, precipitation products; SN, 1/10 of nonprecipitated material. Ribonuclease (RNase) treatment and positions of in vitro translated products are also indicated. (**B**) Pull-down experiments as in (A), using GST fusions of DEK (residues 1 to 100) and the indicated mutant derivatives. CIP indicates treatment with calf intestine alkaline phosphatase.

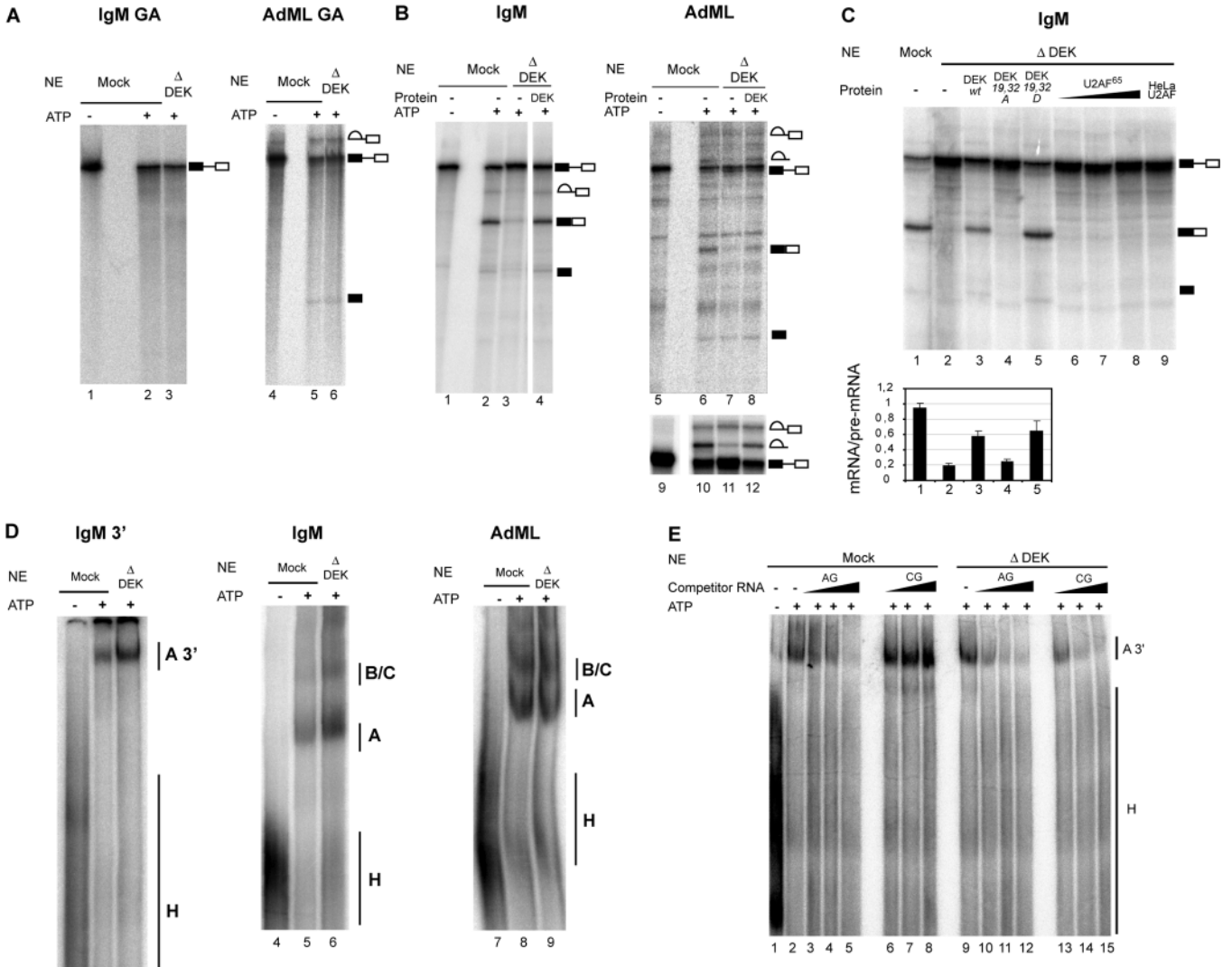



**Fig. 4.** DEK is required for pre-mRNA splicing but not for spliceosome assembly. (**A**) In vitro splicing assays using mouse immunoglobulin IgM or adenovirus major late (AdML) promoter pre-mRNAs in which the 3′ss AG was mutated to GA, in mock- or DEK-depleted nuclear extracts. (**B**) Splicing assays as in (A) using wild-type pre-mRNAs, in the absence or presence of recombinant purified DEK (30 ng/μl). The positions of pre-mRNA, splicing intermediates, and products are indicated. AdML (bottom) shows a clearer effect on lariat intermediates and products. (**C**) Splicing complementation assays as in (B), recombinant DEK; DEK Ala (A) and Asp (D) mutants at serines 19 and 32; recombinant U2AF$^{65}$ (5, 10, and 15 ng/μl); and purified HeLa U2AF heterodimer (10 ng/μl). Quantification of splicing efficiency from three independent experiments is shown. (**D**) Spliceosome assembly corresponding to splicing assays as in (B). IgM 3′, RNA containing 3′ half intron and exon 2 of IgM pre-mRNA. Positions of heterogeneous nuclear RNP (H) and spliceosomal (A and B/C) complexes are indicated. (**E**) Spliceosome assembly assays as in (D) using IgM 3′ RNA in the presence of 50, 200, and 400 molar excess of unlabeled RNAs containing AG or CG at the 3′ss as in Fig. 1B.

dependent adenosine triphosphatase (ATPase) Prp16 with the spliceosome as a timing device that enforces kinetic proofreading of lariat intermediates before the second catalytic step (*11*). An RNA-dependent ATPase could facilitate DEK-mediated displacement of U2AF from pyrimidine tracts not followed by AG 3′ss, as reported for other RNA-protein interactions (*12*). Kinetic proofreading could explain why catalytic activation of the spliceosome requires discrimination between bona fide and cryptic 3′ss during early splicing complex formation. Phosphorylation affects DEK activity in 3′ss recognition and splicing, which suggests that casein kinase II (*13*) or other kinases can influence the selectivity of 3′ss identification and possibly alternative splicing. Phosphorylation could also cause a switch in DEK partners, including those that mediate its association with chromatin or the transcriptional machinery (*7*); this may underlie aspects of the coupling between transcription and pre-mRNA processing (*14*).

### References and Notes


1. P. D. Zamore, J. G. Patton, M. R. Green, *Nature* **355**, 609 (1992).
2. C. L. Kielkopf, N. A. Rodionova, M. R. Green, S. K. Burley, *Cell* **106**, 595 (2001).
3. S. Wu, C. M. Romfo, T. W. Nilsen, M. R. Green, *Nature* **402**, 832 (1999).
4. D. A. Zorio, T. Blumenthal, *Nature* **402**, 835 (1999).
5. L. Merendino, S. Guth, D. Bilbao, C. Martinez, J. Valcárcel, *Nature* **402**, 838 (1999).
6. P. D. Zamore, M. R. Green, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 9243 (1989).
7. T. Waldmann, I. Scholten, F. Kappes, G. Hu, R. Knippers, *Gene* **343**, 1 (2004).
8. T. McGarvey *et al.*, *J. Cell Biol.* **150**, 309 (2000).
9. M. Fornerod *et al.*, *Oncogene* **10**, 1739 (1995) .
10. D. Scherly, W. Boelens, N. A. Dathan, W. J. van Venrooij, I. W. Mattaj, *Nature* **345**, 502 (1990).
11. J. P. Staley, C. Guthrie, *Cell* **92**, 315 (1998).
12. E. Jankowsky *et al.*, *Science* **291**, 121 (2001).
13. F. Kapees *et al.*, *Mol. Cell. Biol.* **24**, 6011 (2004).
14. T. Maniatis, R. Reed, *Nature* **416**, 499 (2002).


15. We thank T. Blumenthal, J. Cáceres, M. Green, T. Nilsen, R. Singh, and Centro de Regulació Genòmica colleagues for comments on the manuscript. L.M.M.S.

was supported by a Praxis fellowship (Portuguese Foundation of Science and Technology, Gulbenkian Ph.D. Program in Biomedicine). Work supported by grants from EU (J.V., M.S.), Ministerio de Ciencia y Tecnología (Spain), and Bundesministerium für Bildung und Forschung (Federal Ministry of Education and Research) (Germany). J.V. dedicates this paper to J. Ortín and A. Mas-Colell.

**Supporting Online Material**
www.sciencemag.org/cgi/content/full/312/5782/1961/DC1
Materials and Methods
Figs. S1 to S9
References and Notes

12 April 2006; accepted 11 May 2006
10.1126/science.1128659

# The Ant Odometer: Stepping on Stilts and Stumps

**Matthias Wittlinger,[1]* Rüdiger Wehner,[2] Harald Wolf[1]**

**Desert ants, *Cataglyphis*, navigate in their vast desert habitat by path integration. They continuously integrate directions steered (as determined by their celestial compass) and distances traveled, gauged by as-yet-unknown mechanisms. Here we test the hypothesis that navigating ants measure distances traveled by using some kind of step integrator, or "step counter." We manipulated the lengths of the legs and, hence, the stride lengths, in freely walking ants. Animals with elongated ("stilts") or shortened legs ("stumps") take larger or shorter strides, respectively, and concomitantly misgauge travel distance. Travel distance is overestimated by experimental animals walking on stilts and underestimated by animals walking on stumps.**

Foraging Saharan desert ants, *Cataglyphis fortis,* use a mode of dead reckoning known as path integration (*1*, *2*) to monitor their current position relative to the nest and to find their way home (*3*). This enables them to return on a direct route, rather than retracing the tortuous outbound journey performed when searching for food items in their flat desert habitat, which is often completely devoid of landmarks. The path integrator requires two kinds of input information: about directions steered, as obtained via the ant's celestial compass (*4*), and about distance traveled, as gauged by the ant's odometer.

The cues by which ants measure travel distance during locomotion have not yet been discovered. However, there are several promising hypotheses. The "energy hypothesis" posits that the (surplus) energy required for locomotion (as opposed to rest) is used to calculate travel distance. This hypothesis is of long standing in arthropod research (*5*), but is not applicable to the problem of the ant odometer, because ants assess their walking distances with great accuracy, irrespective of the load they carry (*6*). The "optic flow hypothesis" has been proven in honeybees, which integrate visual flow-field cues during their foraging flights to gauge flight distance (*7*, *8*). In *Cataglyphis* ants, this mechanism plays a minor role, if it has any effect (*9*) (for further discussion of the optic flow hypothesis, see SOM). Even in complete darkness, on featureless platforms, or with the ventral halves of their eyes covered, the animals are still able to assess traveling distance correctly during their homing runs (*9*, *10*). Also, lateral optic flow does not have any influence at all on distance estimation (*11*). Considering the relatively constant locomotor speed of desert ants, a time-lapse integrator might function to measure walking distance—although this possibility has been refuted in slightly different experimental contexts (*12*). Thus, ants appear to rely primarily on idiothetic cues, most probably derived from the movements of their legs (*13*). Although this step integrator or "pedometer" hypothesis was initially proposed as early as 1904 (*14*), it has remained untested. Here, we examine whether or not ants with manipulated leg lengths, walking on stilts or on stumps, exhibit changes in their stride lengths and, consequently, misgauge their travel distance during homebound runs.

The ants were trained to walk from their nest entrance to a feeder, over a distance of 10 m and in a linear alloy channel (Fig. 1, top). After at least 1 day of training, the animals were caught at the feeding site and transferred to a test channel, aligned parallel to the training channel (Fig. 1, bottom). Once transferred into this test channel, the ants performed their homebound runs, and we recorded the point at which the ants switched from their straight and steady return path to their characteristic nest-searching behavior. This point is marked by a 180° U-turn (*15*), followed by a run pacing back and forth around the anticipated location of the nest entrance. Details of the experimental procedure are given in (*16*).

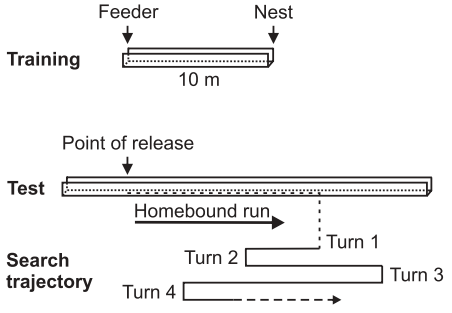



**Fig. 1.** Schematic diagram of channel layout, as well as training and testing procedures applied in *Cataglyphis* foragers (not drawn to scale).

Ants that had reached the feeder on a foraging trip through the training channel were caught and subjected to experimental manipulation. To increase stride length on the animals' homebound runs, their legs were splinted and extended with pig bristles glued to the tibia and tarsus, as illustrated in Fig. 2 (stilts). To decrease stride length, the legs were shortened by severing the tarsomeres in the middle of the tibia segment (Fig. 2, stumps). Operated animals were supplied with a food item and transferred to the test channel, with the food item in their mandibles ("test 1"). These ants started determined homeward runs, and when they had covered the assumed distance to the nest, they switched to the nest-searching behavior outlined above. The ants walking back homeward on stilts clearly overshot [15.30 m, interquartile range (IQR) = 3.24 m, $n$ = 25], whereas ants with shortened legs undershot (5.75 m, IQR = 1.81 m, $n$ = 25) with regard to their normal homing distance (10.20 m, IQR = 2.40 m, $n$ = 25) (Fig. 3A). There are statistically significant

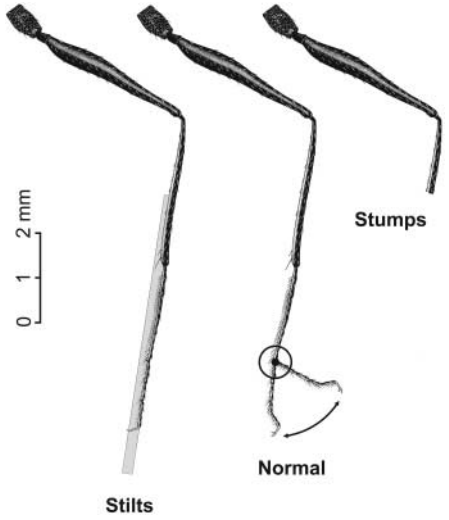



**Fig. 2.** Manipulation of ant legs as performed in the present study. In stilts, attached pig bristles elongated the legs; second, normal unmodified legs, with approximate range of tarsus movement indicated; third, shortened legs (stumps). The right hind leg is shown from anterior [see figure 1 in (*18*)].

[1]Department of Neurobiology, University of Ulm, D-89069 Ulm, Germany. [2]Institute of Zoology, University of Zürich, CH-8057 Zürich, Switzerland.

*To whom correspondence should be addressed. E-mail: matthias.wittlinger@gmx.de

differences ($P \leq 0.001$, Kruskal-Wallis test) among these groups.

The modified and tested ants were put back into their nest and were tested again when they turned up at the feeder during one of the following days ("test 2"). Having already performed their outbound runs to the feeder on stilts or stumps, these animals exhibited homing runs almost identical to those of normal, unmodified ants (Fig. 3B). Ants walking on stilts (stilts, $n = 25$) searched for the nest at 10.55 m (IQR = 1.45 m) distance from the release point and ants with shortened legs at 10.25 m (IQR = 1.76 m) (stumps, $n = 25$). No significant differences were observed between these groups or with regard to unmodified animals (above, 10.20 m).

These results support the hypothesis that desert ants use a pedometer for distance measurement, or a step integrator [loosely speaking, a step counter, although the ants most probably do not literally count (*17*)]. According to the pedometer hypothesis, ants that have traveled to the feeder on normal legs and have had their leg length modified at the feeder should cover a different distance on their homebound journey. This is because the same number of strides made during the outbound travel, as registered by the step integrator, should presumably carry them over longer (stilts) or shorter (stumps) distances, respectively. And later, on reemerging from the nest with their legs still modified, the number of strides will be the same again for outbound and inbound travel (although different from the situation with normal legs in both cases). No such straightforward interpretation is possible for any of the competing hypotheses.

This interpretation rests on the assumption, however, that stride length is indeed altered by the manipulation of leg length, roughly in proportion to the imposed length changes. Thus, a quantitative assessment of the (manipulated) ants' stride lengths and their relation to walking distances is also necessary. It was impossible to count the ants' strides during experiments (for instance, by filming the complete outbound and homebound journeys on high-speed video). Instead, stride lengths were determined in a different set of animals, both normal and manipulated, in a small pen erected on the desert floor. As in most animals, stride length in normal *Cataglyphis* depends on at least two parameters. The first is leg length and, thus, body size, as body morphology is isomorphic in *Cataglyphis fortis* workers (*18*). That is, larger ants take correspondingly longer strides. The second is walking speed. The faster an ant runs, the larger are the strides it takes (*19*–*21*). Normalization of homebound travel distances with regard to both body size and walking speed is thus required for a quantitative interpretation of the above experiments. High-speed video recordings of running ants were analyzed to determine the actual changes in stride lengths caused by the stilts and stumps manipulations described above. Walking speeds were determined in the above experimental animals by the time required for uninterrupted straight segments of homebound travel (and confirmed on the high-speed movies in a different set of animals).

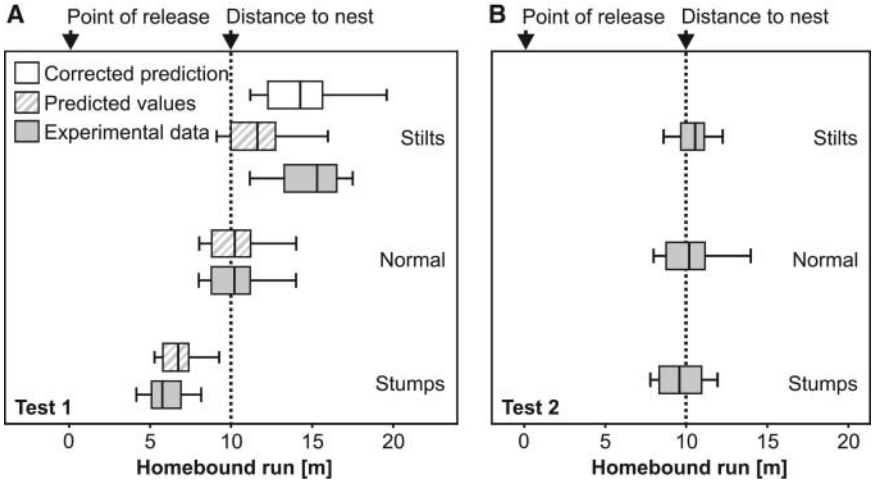



**Fig. 3.** Homing distances of experimental ants, tested immediately after the lengths of their legs had been modified at the feeding site. (**A**) Leg lengths were normal during the outbound journey but manipulated during the homebound run, resulting in different homing distances. (**B**) Ants tested after reemerging from the nest after previous manipulation. In this situation, leg lengths were equal, although manipulated, during outbound and homebound runs. Box plots show median values of the homing distances recorded in $n = 25$ ants per experiment (as well as IQRs, box margins, and 5th and 95th percentiles, whiskers). Median values of the initial six turning points of an ant's nest-search behavior were considered as the centers of search, indicating homing distance. The hatched box plots in (**A**) illustrate the centers of search as predicted from the high-speed video analyses of stride lengths in normal and manipulated animals. The open box represents the prediction corrected for slow walking speed. Details in text.

To derive a quantitative prediction of how far the manipulated ants would have to run on their homebound journey, based on the pedometer hypothesis, the experimental data were first normalized with regard to body size. That is, relative stride lengths with regard to body dimensions were calculated to eliminate effects of size variation. Second, the actual stride lengths were determined in normal and manipulated ants. Ants walking on stilts took significantly longer strides ($P < 0.01$, $t$ test; 14.8 mm, SD = ± 2.5 mm, or +13.9%) than did normal animals (13.0 mm, SD = ± 1.98 mm), and ants on stumps made significantly shorter strides ($P < 0.001$; 8.6 mm, SD = ± 1.73 mm, or –33.2%). These values were then used to derive a prediction for the respective undershoot or overshoot of manipulated ants, based on the assumption that distance is measured by a step integrator, that is, that the ants took the same number of strides during their outbound journey to the feeder and when returning to the nest on manipulated legs (hatched boxes in Fig. 3A). The predicted values show that there is general agreement between the predicted and observed homebound travel distances in manipulated ants, further corroborating the step integrator hypothesis.

However, there are some differences (significant in stilts, $P = 0.015$; not significant in stumps, $P = 0.125$; Mann-Whitney rank sum test) between the observed and the predicted homebound travel distances. Impairment of the ants by the preparation procedures cannot account for these effects. Any serious impairment would tend to shorten the ants' homebound runs. The animals walking on stilts clearly traveled for much longer distances than the normal controls, however, and they did so with apparently normal vigor. Indeed, experimentally modified ants were observed to stilt or stump through their habitat on successful foraging trips several times a day and for many days thereafter (see movie S1 in SOM).

The differences between observed and predicted homing distances may be attributable, though, to altered walking speeds in the manipulated ants. In fact, normal ants traveled at an average 0.31 m/s, and ants walking on stumps averaged 0.14 m/s, a value in good agreement with their shortened legs and stride lengths (see similarity of predicted and experimentally determined homing distances in Fig. 3A). Ants on stilts however, rather than walking at increased speeds, were also slightly slower than normal animals, walking at an average 0.29 m/s. This was presumably due to the added load of glue and pig bristles on their legs. As noted above, changed walking speeds are associated with correspondingly altered stride lengths, which would appear to explain the unexpectedly short prediction value in Fig. 3A (top hatched box). We thus corrected this prediction with the established relations between walking speed and stride length (*19*, *20*). We assumed as a first approximation (and conservatively, when con-

sidering the actually imposed changes in leg length) that ants might run by as much faster on stilts as they ran slower on stumps (0.48 m/s, a value regularly observed in highly motivated normal ants and almost reached by the fastest ants on stilts). This procedure indeed yields a value that is not significantly different from the observed homing distances in ants on stilts (open box in Fig. 3, A; 14.25 m, IQR = 3.35 m), thus confirming the consistency of our data with the step integrator hypothesis.

The slower speeds of the ants walking on stilts further rule out the only alternative explanation of our homing distance data (Fig. 3A, solid boxes). In principle, a step integrator and a time-lapse integrator would both yield the same homing distances, even in ants with manipulated leg and stride lengths, if only the ants kept their stride frequencies constant [or in normal ants, walking speed—which in fact they almost do under normal conditions (*19*, *20*)]. Constant stride frequency would result in a change in walking speed in proportion to altered stride length and a resulting difference in homing distance during a set (outbound) travel time. This assumption is evidently not correct, though, given the walking speeds of the experimental animals.

Future studies will have to address the mechanism of the proposed step integrator, for example, whether it actually registers steps by means of proprioreceptors, or whether it integrates activity of a walking pattern generator, and to what extent sensory feedback regarding stride length and walking performance is considered.

### References and Notes


1. R. Wehner, M. V. Srinivasan, *J. Comp. Physiol. [A]* **142**, 315 (1981).
2. M. L. Mittelstaedt, H. Mittelstaedt, *Naturwissenschaften* **67**, 566 (1980).
3. M. Müller, R. Wehner, *Proc. Natl. Acad. Sci. U.S.A.* **85**, 5287 (1988).
4. R. Wehner, B. Lafranconi, *Nature* **293**, 731 (1981).
5. H. Heran, L. Wanke, *Z. Vergl. Physiol.* **34**, 383 (1952).
6. R. Wehner, in *Animal Homing*, F. Papi, Ed. (Chapman and Hall, London, 1992), pp. 45–144.
7. H. E. Esch, J. E. Burns, *Naturwissenschaften* **82**, 38 (1995).
8. M. V. Srinivasan, S. Zhang, M. Altwein, J. Tautz, *Science* **287**, 851 (2000).
9. B. Ronacher, R. Wehner, *J. Comp. Physiol. [A]* **177**, 21 (1995).
10. M. Thiélin-Bescond, G. Beugnon, *Naturwissenschaften* **92**, 193 (2005).
11. B. Ronacher, K. Gallizzi, S. Wohlgemuth, R. Wehner, *J. Exp. Biol.* **203**, 1113 (2000).
12. S. Wohlgemuth, B. Ronacher, R. Wehner, *Nature* **411**, 795 (2001).
13. H. Mittelstaedt, M. L. Mittelstaedt, *Fortschr. Zool.* **21**, 46 (1973).
14. H. Pieron, *Bull. Inst. Gen. Psychol.* **4**, 168 (1904).
15. S. Sommer, R. Wehner, *J. Comp. Physiol. [A]* **190**, 1 (2004).
16. Materials and methods are available on *Science* Online.
17. N. R. Franks *et al.*, *Proc. R. Soc. London B. Biol. Sci.* **273**, 165 (2006).
18. R. Wehner, *Senckenbergiana Biol.* **64**, 89 (1983).
19. C. P. E. Zollikofer, thesis, University of Zürich (1988).
20. C. P. E. Zollikofer, *J. Exp. Biol.* **192**, 95 (1994).
21. C. P. E. Zollikofer, *J. Exp. Biol.* **192**, 107 (1994).


22. Funded by the Volkswagen Stiftung (I/78 580 to H.W. and R.W.), the Swiss National Science Foundation (3100-61844 to R.W.), and the Universities of Ulm and Zürich.


### Supporting Online Material

www.sciencemag.org/cgi/content/full/312/5782/1965/DC1
Materials and Methods
SOM Text
References and Notes
Movie S1

2 March 2006; accepted 26 May 2006
10.1126/science.1126912

# Social Modulation of Pain as Evidence for Empathy in Mice

**Dale J. Langford, Sara E. Crager, Zarrar Shehzad, Shad B. Smith, Susana G. Sotocinal, Jeremy S. Levenstadt, Mona Lisa Chanda, Daniel J. Levitin, Jeffrey S. Mogil***

**Empathy is thought to be unique to higher primates, possibly to humans alone. We report the modulation of pain sensitivity in mice produced solely by exposure to their cagemates, but not to strangers, in pain. Mice tested in dyads and given an identical noxious stimulus displayed increased pain behaviors with statistically greater co-occurrence, effects dependent on visual observation. When familiar mice were given noxious stimuli of different intensities, their pain behavior was influenced by their neighbor's status bidirectionally. Finally, observation of a cagemate in pain altered pain sensitivity of an entirely different modality, suggesting that nociceptive mechanisms in general are sensitized.**

Although most consider true empathy to be an exclusive ability of higher primates, empathy may be a phylogenetically continuous phenomenon with subclasses such as "emotional contagion" well within the reach of all mammals (*1*). However, there is little evidence for adult-adult empathy outside of primates. In rats (*2*) and pigeons (*3*), the pain-related distress of a conspecific can serve as a conditioning stimulus. Rats produced operant responses to terminate the distress of a conspecific (*4*), but this might be better explained by arousal than altruism (*5*). One theory of human empathy postulates "physiological linkage" between empathizing individuals (*6*). In one study, empathic accuracy for negative emotion was highest in those dyads featuring high levels of time synchrony of autonomic measures (*7*). We hypothesized that if empathy does indeed exist in mice, the real-time observation of pain in one mouse might affect the responses of its conspecifics to painful stimuli.

We first used a sensitive nociceptive assay, the reflexive 0.9% acetic acid abdominal constriction ("writhing") test. We placed mice singly within transparent Plexiglas cylinders to observe writhing behavior. For comparison, we placed two same-sex mice within each cylinder and injected either one or both mice. In the "both writhing" (BW) condition, each mouse observed the other in pain; in the "one writhing" (OW) condition, the injected mouse observed an uninjected counterpart. BW mice displayed significantly more pain behavior than isolated mice, but only when their counterparts were cagemates (Fig. 1A). The hyperalgesia was marginally enhanced in same-sex siblings living together, but a separate experiment confirmed that close genetic relatedness was not required (fig. S1). Writhing behavior in BW dyads co-occurred in time at levels significantly exceeding those expected by chance (Fig. 1B) and significantly more so in cagemate pairs than stranger pairs. The hyperalgesia and behavior co-occurrence developed over 14 to 21 days of being housed together (Fig. 1, C and D). In general, observed behaviors other than writhing were similar across all conditions (figs. S2 and S3), although evidence suggested higher levels of anxiety or stress produced by the noxious stimulus in stranger pairs relative to cagemates (fig. S4). Because the observed effects on pain behavior were higher in cagemates, stress is not a likely mediator.

When strangers were tested in dyads, a significant decrease in writhing behavior was observed in the OW condition compared to that observed in isolation (Fig. 1A). The inhibition was entirely specific to males (fig. S5) and is likely due to distraction or social stress–induced analgesia.

These findings imply the communication of pain from one mouse to another. To determine the transmitting sensory modality, we blocked sensory inputs individually, by placing physical barriers to sight and/or touch or by rendering mice anosmic or deaf (*8*). The only manipulation that significantly abolished the BW/OW hyperalgesia was a visual blockade using an opaque Plexiglas barrier (Fig. 2A). [Despite their albinism, the CD-1 mice used in these studies display no deficits in visually dependent behavioral tasks (*9*).] The opaque barrier also

Department of Psychology and Centre for Research on Pain, McGill University, Montreal, QC H3A 1B1, Canada.

*To whom correspondence should be addressed. E-mail: jeffrey.mogil@mcgill.ca

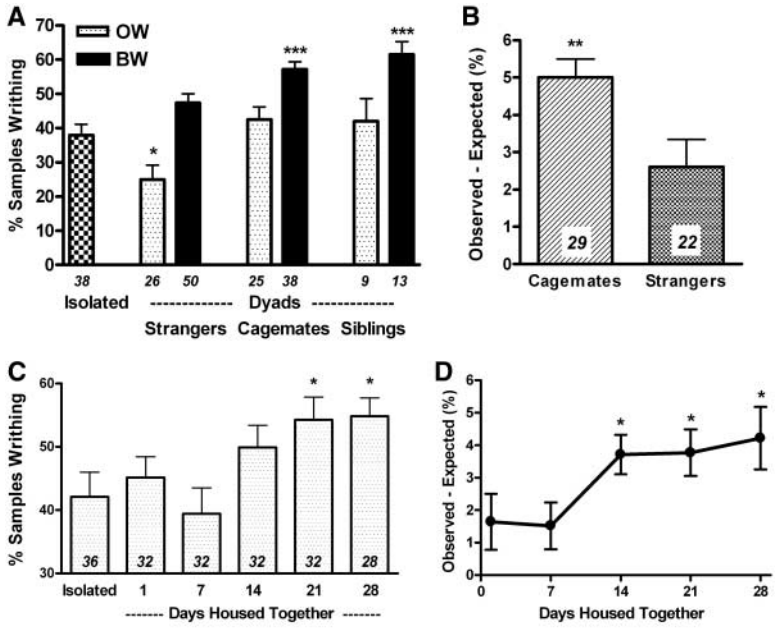



**Fig. 1.** (**A** to **D**) Mice injected with 0.9% acetic acid in the presence of similarly injected cagemates display higher levels of pain behavior, which co-occurs in time. In all graphs, group sample sizes are indicated in italics. (A) Mice were tested in isolation (Isolated), or in dyads where either one mouse (One Writhing; OW) or both mice (Both Writhing; BW) received acetic acid injections. Bars represent the mean ± SEM percentage of sampled intervals showing writhing behavior (% Samples Writhing). *$P < 0.05$, ***$P < 0.005$ by Dunnett two-way case-control comparison posthoc test compared to Isolated mice. (B) Statistically significant co-occurrence in writhing behavior in the Cagemates and Strangers conditions (sign test, $P < 0.05$ in both cases); the co-occurrence was significantly higher in Cagemates. Using data from (A), the expected number of samples with writhing in both mice of the dyad was calculated as a joint probability. Bars represent the mean ± SEM excess of observed samples with joint writhing above the expected value, as a percentage. **$P < 0.01$ compared to Strangers (Student's $t$ test). (C) Data from a separate experiment using naïve mice housed together for 1, 7, 14, 21, or 28 days and tested in BW dyads. Isolated mice were taken from the 28-day group, but were tested alone. Bars are as in (A). *$P < 0.05$ by Dunnett one-way case-control comparison posthoc test compared to Isolated mice. Data in (D) were calculated from subjects shown in (C); symbols represent the mean ± SEM excess of observed samples with joint writhing above the expected value, as a percentage. *$P < 0.05$ compared to zero (sign test). Significant linear trends were evinced in (C) and (D) ($P = 0.001$ and $P < 0.005$, respectively).

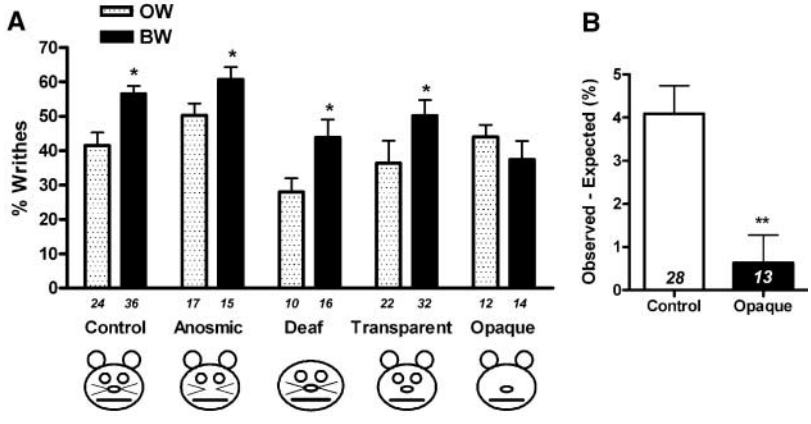



**Fig. 2.** (**A** and **B**) Apparent dependence of socially mediated hyperalgesia and co-occurrence on visual cues. Mice, all cagemates ($n$ = 10 to 36 per group; housed together for >21 days), were tested in dyads as described in Fig. 1, such that either one mouse (One Writhing; OW) or both mice (Both Writhing; BW) received 0.9% acetic acid injections. "Control" data (intact mouse face cartoon) were taken from Cagemates condition in Fig. 1 for purposes of comparison. (A) Bars represent the mean ± SEM percentage of sampled intervals showing writhing behavior (% Samples Writhing). *$P < 0.05$ by Student's $t$ test compared to OW group. The significantly lower writhing behavior of the Deaf-OW group reflects the relative insensitivity to the noxious stimulus of the BALB/c strain, as previously reported (*24*). (B) The abolition of writhing behavior co-occurrence in BW dyads in which one mouse is prevented from seeing the other (Opaque condition). Bars represent the mean ± SEM excess of observed samples with joint writhing above the expected value, as a percentage. **$P < 0.01$ compared to Control group (Student's $t$ test).

blocked the co-occurrence of writhing behavior in the BW condition (Fig. 2B). Zinc sulfate treatment destroys the olfactory epithelium in the mouse but spares axonal transport from the vomeronasal organ to the accessory olfactory bulb (*10*), and thus pheromonal communication cannot be ruled out. It is, of course, highly likely that the recognition of the other mouse in the dyad as stranger, familiar, or sibling was achieved via olfactory cues (*11*), which were likely unimpeded by the barriers. Indeed, social communication is recognized to be commonly multimodal in many species (*12*).

An existing data set (*13*) provided an independent verification of the social co-occurrence of pain behavior in simultaneously tested mice, in another assay. In the 5% formalin test, licking behavior was statistically time-synchronized within runs of four mice tested individually in Plexiglas observation cylinders, but in close proximity and in full view of each other (figs. S6 and S7A). The co-occurrence of pain behaviors in familiar individuals may itself be evidence of empathy, representing a compelling analog to the demonstrations of physiological linkage in empathizing humans (*7*).

These formalin data also showed a reduction of between-subject variance within a run (fig. S7B), suggesting that subjects' pain behaviors were being influenced, perhaps bidirectionally, by their neighbors. In a new experiment, we compared pain behavior in "both licking" dyads in which both mice received either a high dose (5%) or a low dose of formalin (1%), or in which each mouse received different doses (1%, 5%). Pain behavior was influenced by that of the neighbor mouse, such that licking times were marginally increased in mice receiving the low dose while observing a high dose–injected cagemate, and significantly reduced in mice receiving the high dose while observing a low dose–injected cagemate (Fig. 3). No significant effects were observed among strangers (fig. S9).

Finally, we investigated whether the observation of a cagemate in pain could modulate sensitivity to pain of a wholly different modality. Mice were tested in dyads as described, but in addition to measuring writhing behavior, we tested all mice for their sensitivity to withdraw from a noxious radiant heat stimulus before and at 5-min intervals after injection of acetic acid (or no injection). Injection and the mere observation of a cagemate's writhing behavior both produced significant and equivalent thermal hyperalgesia (Fig. 4). No observation effects whatsoever were observed among strangers (fig. S10). Concurrent thermal pain testing did not abolish the BW/OW increase in writhing behavior (Fig. 4C), and a significant correlation was observed between the writhing behavior of

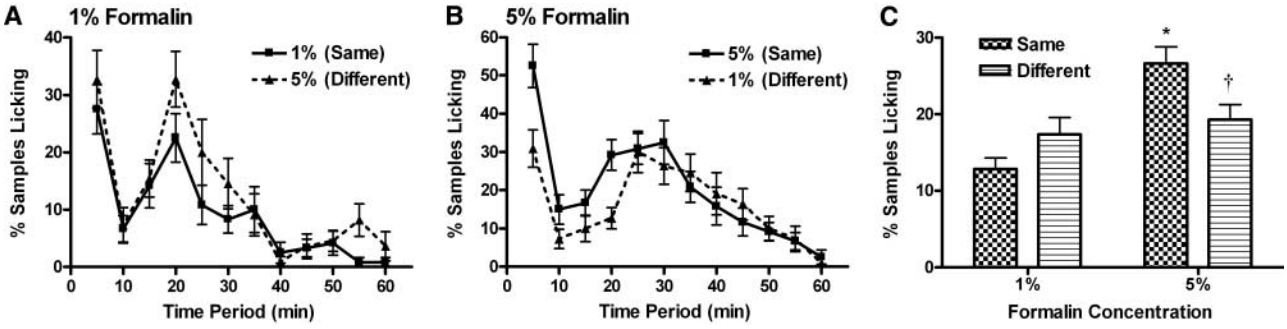



**Fig. 3.** (**A** and **B**) Bidirectional modulation of pain behavior produced by observation of a cagemate in the formalin test. Mice, all nonsibling cagemates ($n$ = 22 to 24 per group; housed together for >21 days), were tested in dyads. In the "Same" condition, both mice received either 1% formalin or 5% formalin. In the "Different" condition one mouse received 1% formalin and the other received 5% formalin. All groups displayed the expected biphasic pattern of responding (A and B). A two-way (injected dose × observed dose) repeated measures analysis of variance (ANOVA) revealed a significant three-way interaction ($P < 0.05$). (A) Data from all mice receiving 1% formalin; the legend describes the status of the other mouse in the dyad. (B) Data from all mice receiving 5% formalin; the legend describes the status of the other mouse in the dyad. In (A) and (B) (note different ordinate scales), symbols represent the mean ± SEM percentage of sampled intervals showing formalin-induced recuperative behavior (% Samples Licking) per 5-min time bin. (**C**) Totals in all conditions from 0 to 40 min after injection, after which there was no longer significantly different licking behavior between 1% and 5% groups. ANOVA revealed a highly significant injected dose × observed dose interaction ($F_{1,88} = 9.3$, $P < 0.005$). *$P < 0.05$ compared to analogous 1% condition. †$P < 0.05$ compared to analogous "Same" condition.

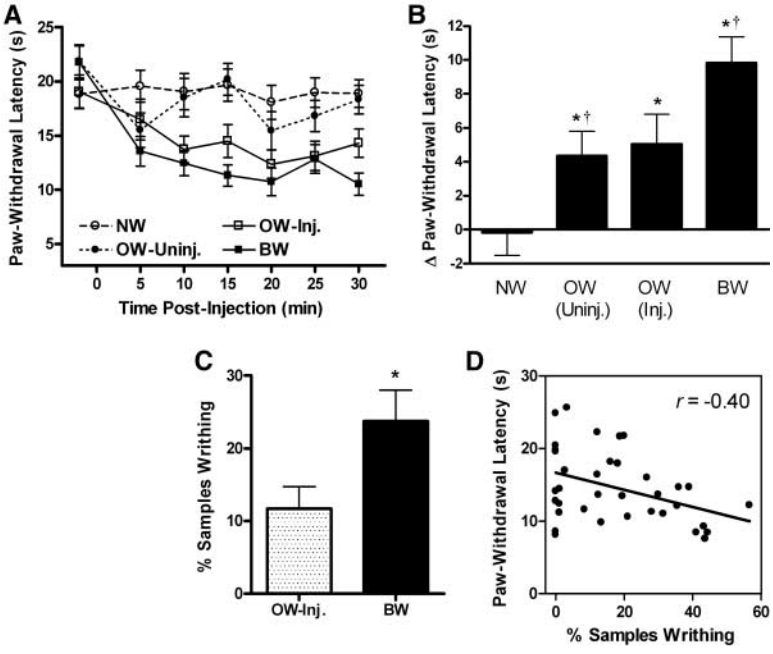



**Fig. 4.** (**A** to **D**) Thermal hyperalgesia produced by injection of acetic acid, by mere observation of a cagemate injected with acetic acid, or both. Mice (all nonsibling cagemates; $n$ = 28 to 31 per group; housed together for >21 days) were tested in dyads as described in Fig. 1. Before injection, all mice were tested for baseline thermal sensitivity. In the BW ("both writhing") group, both mice were removed at time = 0, given an injection of 0.9% acetic acid, and returned to their cylinder. In the NW ("none writhing") group, both mice were removed and replaced, with neither receiving any injection. In the OW ("one writhing") group, one mouse received an acetic acid injection (OW-Inj.) and the other (OW-Uninj.) did not. All mice were retested for thermal sensitivity at 5-min intervals for 30 min. Symbols in (A) represent the mean ± SEM paw-withdrawal latencies (average of both hindpaws). Bars in (B) represent the mean ± SEM average change in paw-withdrawal latencies from the baseline latency. *$P < 0.05$ compared to NW group and zero; †$P < 0.05$ compared to the group immediately to the left. Bars in (C) represent the mean ± SEM percentage of sampled intervals showing writhing behavior (% Samples Writhing) of mice receiving acetic acid (both mice in BW group; OW-Inj. mice). *$P < 0.05$ compared to OW-Inj. group. (D) A significant correlation ($r = -0.40$; $P < 0.05$) between the writhing behavior of one mouse in a dyad (ordinate; BW and OW-Inj. only) and the average (postinjection) paw-withdrawal latency of its dyadic counterpart (abscissa; BW and OW-Uninj. only).

one mouse in the dyad and the thermal hyperalgesia exhibited by the other (Fig. 4D). These data suggest that the pain system is sensitized in a general manner by the observation of pain in a familiar, and furthermore demonstrate that socially mediated hyperalgesia can be elicited in the clear absence of imitation. Mechanisms underlying these phenomena are thus more likely to be found in the sensory/perceptual system than in the motor system.

Rodents are known to recognize and have emotional reactions to the pain of conspecifics (*2*), and their pain sensitivity can be altered by social factors (*14*–*17*). However, most of these studies reported analgesia rather than hyperalgesia and did not evaluate effects in real time, when another's pain was actually being observed. These phenomena may represent an example of coaction social facilitation, depending on one's definition of that term (*18*). However, our findings are consistent with the perception-action model of empathy proposed by Preston and de Waal (*1*), both in the automatic priming of somatic responses in a state similar to that of the attended object and in the modulating effects of familiarity and similarity of experience between subject and object. Our observations cannot be easily explained by stress, imitation, or conditioning, and they neither depend on nor necessarily indicate the presence of sympathy, conscious (cognitive) representations, or altruism. Empathy for pain is currently a topic of much study in humans (*19*–*21*), and "mirror neurons" responding to another's pain may have been identified in human anterior cingulate cortex (*22*). A large human literature documents the effects on pain report of observation of pain in others (*23*); the present data suggest that these effects may be mediated precognitively. There are clear limitations to the mechanistic information that can be gleaned from human

studies; the availability of an animal model of empathy will allow the application of far more powerful experimental techniques.

### References and Notes


1. S. D. Preston, F. B. M. de Waal, *Behav. Brain Sci.* **25**, 1 (2002).
2. R. M. Church, *J. Comp. Physiol. Psychol.* **52**, 132 (1959).
3. S. Watanabe, K. Ono, *Behav. Processes* **13**, 269 (1986).
4. G. E. Rice, P. Gainer, *J. Comp. Physiol. Psychol.* **55**, 123 (1962).
5. J. J. Lavery, P. J. Foley, *Science* **140**, 172 (1963).
6. H. B. Kaplan, S. W. Bloom, *J. Nerv. Ment. Dis.* **131**, 128 (1960).
7. R. W. Levenson, A. M. Ruef, *J. Pers. Soc. Psychol.* **63**, 234 (1992).
8. Materials and methods are available as supporting material on *Science* Online.
9. B. Adams, T. Fitch, S. Chaney, R. Gerlai, *Behav. Brain Res.* **133**, 351 (2002).
10. K. McBride, B. Slotnick, F. L. Margolis, *Chem. Senses* **28**, 659 (2003).
11. C. Dulac, A. T. Torello, *Nat. Rev. Neurosci.* **4**, 551 (2003).
12. S. Partan, P. Marler, *Science* **283**, 1272 (1999).
13. S. G. Wilson *et al.*, *Pain* **96**, 385 (2002).
14. P. Raber, M. Devor, *Pain* **97**, 139 (2002).
15. M. S. Fanselow, *Behav. Neurosci.* **99**, 589 (1985).
16. G. Agren, K. Uvnas-Moberg, T. Lundeberg, *Neuroreport* **8**, 3073 (1997).
17. F. R. D'Amato, F. Pavone, *Behav. Neural Biol.* **60**, 79 (1993).
18. D. A. Clayton, *Q. Rev. Biol.* **53**, 373 (1978).
19. T. Singer *et al.*, *Science* **303**, 1157 (2004).
20. P. L. Jackson, A. N. Meltzoff, J. Decety, *Neuroimage* **24**, 771 (2005).
21. A. Avenanti, D. Bueti, G. Galati, S. M. Aglioti, *Nat. Neurosci.* **8**, 955 (2005).
22. W. D. Hutchison, K. D. Davis, A. M. Lozano, R. R. Tasker, J. O. Dostrovsky, *Nat. Neurosci.* **2**, 403 (1999).
23. K. D. Craig, S. M. Weiss, *J. Pers. Soc. Psychol.* **19**, 53 (1971).
24. J. S. Mogil *et al.*, *Pain* **80**, 67 (1999).


25. This work was supported by the Louise Edwards Foundation. We thank E. Balaban, C. Bushnell, and J. Lund for helpful discussions.


### Supporting Online Material

www.sciencemag.org/cgi/content/full/312/5782/1967/DC1
Materials and Methods
SOM Text
Figs. S1 to S10
References

4 April 2006; accepted 26 May 2006
10.1126/science.1128322


page 1894

INTRODUCTION

# *Science* Looks at Life

WHAT HAPPENS WHEN RESEARCHERS APPLY THE DISCIPLINE OF SCIENCE TO THE RICH—often chaotic—ferment of our lives? This issue provides a sampling of provocative insights by social scientists studying humans at different stages of the life cycle. We thank Robert Axelrod, William Butz, and Barbara Torrey, who aided us in this project, part of an ongoing program of encouraging the best in the social sciences.

Where do the data come from? Butz and Torrey (p. 1898) introduce the special section by describing new technologies and some of the challenges that remain in their application. Horizons are being opened through use of the Internet, geographical information systems, and biomarkers. Genetics is providing new insights into the etiology of our political views, according to Kinder (p. 1905), who describes evidence that we are hard-wired for some of our political tendencies, although complex changes occur throughout life.

Helping children grow and develop is something that cuts to the heart of every community. Heckman (p. 1900) believes that the United States is not spending its child education money wisely—that we spend too little at critical points in child development before kindergarten and too much in "second-chance" efforts when interventions have become costly and ineffective. Although Richter (p. 1902) also considers early involvement to be crucial, she notes that even children who have lived in poverty and witnessed or experienced violence in South Africa often find whatever opportunities are available to become healthy, caring members of society.

What is important in life, and would we know it if we saw it? Measuring happiness or satisfaction is not the same as measuring height; it changes depending on what we are focusing on at the time (Kahneman *et al.*, p. 1908). Our perceptions also tend to change as we grow older (or have other reasons to feel that the time left to us is shrinking and finite); this can lead us to devote more energy to strengthening current bonds and less to striking out on new adventures (Carstensen, p. 1913). We are seeing a shift in the proportions of young and old in many parts of the world. A News story by Balter (p. 1894) explores declines in human fertility and how fretful governments are responding. Vaupel and Loichinger (p. 1911), in their analysis of changing age demographics in Europe, describe a way to change the balance between family and work that will strike a responsive chord in many working parents.

Without insights from the social sciences, technological breakthroughs may never be translated into real-world solutions, as described in the Editorial by Lane (p. 1847). Social sciences are integrated into many parts of *Science*, in addition to research, as in ScienceCareers.org, which has a portal devoted to social, behavioral, and economic sciences. This week, *Science* Careers is focusing on the decision, risk, and management sciences. A major avenue of scholarship for social scientists is the book, and the Books section (p. 1876) contains reviews of three political science–related offerings. As a further resource, readers should note the Science of Aging Knowledge Environment (SAGE KE), which has for the past 5 years provided a variety of views on aging, including a Perspective this week on changing work patterns (Levine *et al.*).

Rarely has a group of papers been so engaging and stimulating as this foray into the frontiers of the social sciences. We hope the experience will be one of many.

–BARBARA JASNY, DONALD KENNEDY, ELIOT MARSHALL

## Life Cycles

### CONTENTS



*See also related Editorial page 1847; Book Reviews pages 1876, 1877, 1878; SAGE KE and* Science *Careers material on p. 1843 or at www.sciencemag.org/sciext/lifecycles/*

Science





*Published by AAAS*

CREDIT: JOCHEN LUEBKE/AFP/GETTY

NEWS

# The Baby Deficit

**As fertility rates decline across the developed world, governments are offering big incentives for childbearing. Experts don't expect them to have much effect**

Last month, from the podium of the Kremlin's grandiose Marble Hall, Russian President Vladimir Putin expounded on subjects vital to his nation's future—economic growth, technological modernization, and world trade—then he turned to the "most important" matter. "What I want to talk about," Putin said in his annual speech before the Federal Assembly, "is love, women, children. I want to talk about the family, about the most acute problem facing our country today—the demographic problem." Reminding the deputies that Russia's 143-million-strong population was declining by almost 700,000 people each year, Putin proposed a fistful of incentives to boost the country's flagging birthrate. They include raising the childcare benefit of 700 rubles ($26) per month to 1500 rubles for a first child and 3000 rubles for a second child, and paying 18 months of maternity leave equal to at least 40% of a mother's previous wages.

Putin is not the only politician talking about babies these days. Earlier this year, Poland's Parliament approved a one-time payment of 1000 zlotys ($328) for each child born, and this month, German Chancellor Angela Merkel proposed a 1-year paid leave for women who have children. When Australia introduced its own generous "baby bonus" in 2004, the country's treasurer Peter Costello exhorted parents to have "one for Mum, one for Dad, and one for the country." On 1 July, Australia's bonus will jump from $2250 to $3002 per child (in U.S. dollars) and will reach $3762 by 2008. Meanwhile, pro-family inducements have been in place for many years in France, Sweden, and other European countries.

Political leaders and economists see plenty of justification for spending all this money. In the European Union (E.U.), for example, low birthrates have already begun to shrink the population, and demographers project that the E.U. will lose between 24 million and 40 million people during each coming decade unless fertility is markedly raised (*Science*, 28 March 2003, p. 1991). Population losses could bring a raft of negative economic consequences in the industrialized world, as well as greater stresses on social security and health care systems as the proportion of older citizens increases. "The changes projected for the United States are not as dramatic as those projected for other areas—particularly Europe and Japan—but they nonetheless present substantial challenges," then–Federal Reserve Board chair Alan Greenspan told a 2004 symposium on population aging in Jackson Hole, Wyoming.

> **"The popularity of baby-bonus schemes** among governments is difficult to understand."
>
> —Anne Gauthier, University of Calgary

Although these trends are most pronounced in the developed world, fertility declines are now also being detected even in less affluent areas of Latin America and Asia. Roughly half of the world's nations, with more than 40% of the human population, now have birthrates below replacement levels, and fertility rates are falling steadily in most developing countries as well. To be sure, demographers predict that the world's population will continue to increase for decades to come, rising from its current 6.5 billion to somewhere between 8 billion and 11 billion by 2050 (see sidebar, p. 1896). But nearly all of this increase will be in developing countries.

Population researchers nevertheless are currently engaged in a lively debate over just what, if anything, developed countries can do to increase family size. Some believe very low fertility rates are here to stay. "The popularity of baby-bonus schemes among governments is difficult to understand," says Anne Gauthier, a sociologist at the University of Calgary in Canada. "While the additional financial support is bound to be welcomed by parents, the overall effect on fertility is likely to be small."


Published by AAAS

**Doing her part.** Minister for Families Ursula von der Leyen (with her children) oversees Germany's effort to increase fertility rates.

Others argue that even modest boosts in the birthrate can make a difference. "We can only expect relatively small effects of policy on fertility, but relatively small effects are important when fertility is low," says demographer Peter McDonald of the Australian National University (ANU) in Canberra, whose advocacy of pro-family policies helped bring about Australia's baby bonus. Yet both sides agree that falling fertility rates might be irreversible once they drop below a certain level—what some demographers have begun to call the "low-fertility trap."

### The demographic transition

Predicting population trends is a tricky business, fraught with assumptions about what humans are likely to do in the future. Most demographers rely on a complex parameter called the total fertility rate (TFR). For any particular country and year, the TFR is a hypothetical measure of the average number of children that nation's women would bear during their lifetimes if, at each stage of their lives, they behaved exactly like women in each age group did during that year. By comparing TFRs from one year to another, demographers can track fertility trends. Leaving aside the effects of immigration and emigration, if a population is to remain the same size, both parents must replace themselves. For industrialized countries, demographers define a replacement-level TFR as 2.1—slightly more than a flat rate, to account for the small fraction of children who die before reaching reproductive age.

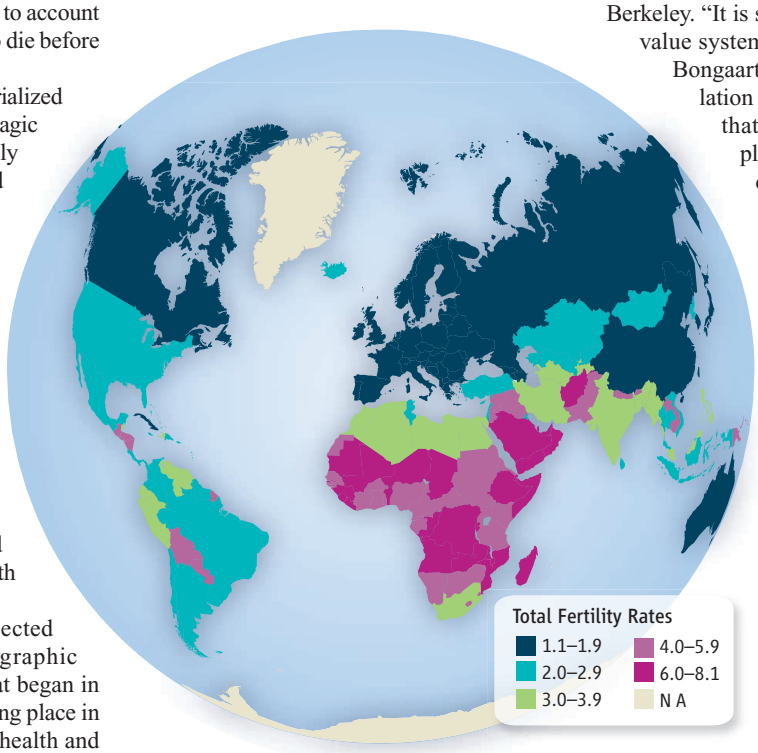
Yet nearly all of the world's industrialized nations have TFRs well below this magic number. Russia's current TFR is only 1.28 (which ties it with Italy and Spain), Poland's is 1.25, Germany's is 1.39, and Australia's is 1.76, which helps explain the alarm expressed by political leaders in those countries. Even the E.U. nations with the highest birthrates, France and Ireland, are falling short of replacement, with TFRs of 1.84 and 1.86, respectively. Nor is the baby shortage restricted to Europe: South Korea's TFR is 1.27 and Japan's is 1.25. Only the United States, exceptional in the developed world, hits the replacement mark, with a TFR of 2.09.

Today's low TFRs are an unexpected consequence of a so-called demographic transition to lower fertility rates that began in Europe in about 1800 and is still taking place in much of the world. As advances in health and hygiene increased the likelihood of a child surviving to reproduce, both death and birthrates started to fall, especially in industrialized countries. Although TFRs remain high in some of the world's poorest countries—Niger has the highest TFR, 7.46—the demographic transition is either under way or completed in most nations. The process has taken place even in relatively poor countries such as Mexico, where TFR dropped from 6.5 to 2.5 between 1975 and 2005, and the Philippines, which saw a decline from 6.0 to 3.2 during the same period. However, demographers had assumed that the decline would stop when replacement-level TFRs were reached. "During the early 1970s, everyone talked about the magic floor of replacement," says David Reher, a population historian at the Complutense University of Madrid, Spain. "Nobody thought it would go below 2.1."



**Proud papa.** Australian Treasurer Peter Costello fathered a baby-bonus scheme.

Yet by 1975, several European countries, as well as the United States and Canada, had already dipped below this floor. (Although the United States has now come back up to replacement level, Canada's TFR has continued to plummet and now stands at 1.61.) This trend, which many demographers and economists call the "second demographic transition," has its roots in the social changes that swept much of the Western world during the 1960s and 1970s. As women entered the labor force in increasing numbers and obtained easier access to effective contraception and as conflicts between work and childbearing intensified, parents began to delay the timing of their first child, which inevitably led to a reduction in the total number of offspring. These shifts were accompanied by a constellation of new attitudes toward family, career, and personal autonomy that are not easily quantified, researchers say. "Human reproductive behavior is profoundly social," says Jennifer Johnson-Hanks, a demographer and anthropologist at the University of California, Berkeley. "It is structured by social categories, value systems, and power relations." John Bongaarts, a demographer at the Population Council in New York City, adds that personal choice has come to play a much bigger role in reproductive decisions. In earlier days, Bongaarts says, "people tended to do what society expected of them. Over time, individual agency has become more important."




Social factors also explain the United States's anomalously high fertility rate, population experts say. Although relatively higher birthrates among some ethnic groups and more recently arrived immigrants, including Hispanics, explain part of the difference, the TFR for non-Hispanic whites is still about 1.85, equivalent to the highest rates seen in Europe. "There are several factors that make the TFR in the U.S. higher than

CREDITS (TOP TO BOTTOM): DAVID CAIRD/*THE SUN HERALD*; SOURCE: U.N. POPULATION PROSPECTS


Published by AAAS

in many European countries," Bongaarts says, including a higher rate of unwanted pregnancies due to restrictions on birth-control information, a lower unemployment rate, and a greater tendency for women to have children earlier in life than in Europe. Gauthier adds that a stronger emphasis on religion and "traditional values" in the United States also tends to favor larger families.

### Aged and dependent

The key reason that economists and other experts are worried about low fertility rates is that they accelerate an overall "aging" of a population, in which the proportion of elderly adults relative to the active labor force increases. The consequences of an increase in this so-called dependency ratio are hard to predict, says demographer James Trussell of Princeton University. "The economic burden of the elderly will depend on their health, on employment opportunities, and on the social institutions that support their care," Trussell says. "But it is clear that it will be a challenge." One way that many developed countries meet the challenge now is through immigration, which tends to increase the number of younger workers. Yet few demographers see immigration as the answer.

"As a short-term solution, it is necessary, and it is happening," says Reher. "But there are very serious doubts about whether it is a long-term solution. Migrant fertility starts higher than that of the native population but very quickly descends towards local fertility levels." Trussell agrees: "To have an appreciable effect on the aging of a population, you would need massive immigration, which is not politically acceptable in either Europe or the U.S."

That leaves raising birthrates as the only solution, assuming that a solution to low fertility rates is possible—and desired. Some demographers take heart in an apparent gap between how many children parents would ideally like to have if they felt they could manage it and how many they actually do have. In this gap, some see wiggle room for fertility-enhancing policies. Thus, public-opinion surveys carried out by the E.U. as part of its Eurobarometer program have suggested that this gap amounts to an average of about 0.5 children per woman. Indeed, baby bonuses and other pro-family measures are in part designed to make it easier financially for families to fulfill this ideal. But Gauthier questions whether the gap is actually that large. In a study in press at the journal *Population Research and Policy Review*, she concludes that the "window of opportunity" for family policies might actually be as little as 0.1 to 0.2 children per woman.

Gauthier and other researchers agree nevertheless that pro-family policies have had some positive effect on fertility rates in countries such as France, whose TFR of 1.84 is the second highest in the E.U. after Ireland. "There are no fewer than 38 measures in favor of families with children," says demographer Laurent Toulemon of the National Institute of Demographic Studies in Paris. For example, mothers receive 16 weeks of maternity leave at more than 80% of their normal pay, which is extended to 26 weeks beginning with a third child. Parents also receive numerous direct allowances to help provide for young children, and the number of publicly funded nursery schools has expanded in recent years to the extent that nearly every child is guaranteed a place. In fact, there are so many pro-family policies, says Toulemon, "that it is almost impossible to evaluate the impact of each one" on fertility.

**Allons les enfants!** France guarantees nursery school spots to nearly all children.

CREDIT: CHARLES JEAN MARC/CORBIS SYGMA

Despite these generous allotments, however, France's relatively high fertility rate in European terms is still below replacement. The same is true of Sweden, where government officials credit bountiful policies designed to make life easier for working parents with recent gains in TFR from about 1.6 to 1.8. Yet Gigi Santow, formerly of Stockholm University and now an independent demographer in Sydney, Australia, says that this fertility jump was not due to baby bonuses or other direct attempts to create a baby boom. "Swedish fertility rates may well have responded to the government's integrated web of cradle-to-grave social policies," Santow says. She adds that fertility plummeted during the economic recession that hit Sweden during the 1990s, despite the policies then in place.

Proving that financial incentives can actually raise fertility rates is very difficult—and demographers do not always agree. "We cannot carry out an experiment," says Gauthier. "We can only look historically at what has happened and rely on cross-national differences in policies." Earlier this month, for example, Australia's news media were abuzz with reports of the latest birth figures from the Australian Bureau of Statistics, show-

## The Bomb That Wasn't

When Stanford University entomologist Paul Ehrlich published *The Population Bomb* in 1968, the world's human population was about 3.5 billion. Today, it is approximately 6.5 billion. Yet the worst of Ehrlich's widely publicized predictions, including the starvation of hundreds of millions of people in mass famines, have not come true. Still, the world's population is expected to continue to grow until at least 2050, according to estimates by the United Nations Population Division (esa.un.org/unpp). Just how much it will increase depends on future fertility, which is very difficult to predict. U.N. population experts have examined three hypothetical fertility trends, which they term medium, low, and high. Under the medium scenario, population would reach 9.1 billion by 2050, but the low and high scenarios project as few as 7.6 billion people and as many as 10.6 billion.

Nearly all of this growth will be in developing countries, with major contributions from nations such as India, Pakistan, Nigeria, Bangladesh, and China. (Even the United States, with its relatively youthful population, will add significant numbers.) Fueled by very high fertility rates, between now and 2050, population is expected to at least triple in some nations, such as Afghanistan, Burundi, Chad, Democratic Republic of Congo, Mali, and Uganda—despite high HIV infection rates in many African countries. Yet over the long term, fertility is expected to drop dramatically in even the poorest countries, from an average of five children per woman now to about 2.6 in 2050; and under the U.N.'s medium scenario, average worldwide fertility will decline to 2.05 by 2050, and to just over 1.5 in the low scenario, well below the replacement level.

"Virtually all countries are headed towards replacement-level fertility or below," says Ronald Lee, a demographer at the University of California, Berkeley. "But there may be pauses and reversals along the way, sometimes lasting decades." If so, the population bomb may ultimately fizzle out—that is, assuming an already stressed planet can survive the onslaught of 9 billion human beings.

**–M.B.**



Published by AAAS

ing that 261,404 babies were born in 2005, 2.4% more than the previous year and the highest number since 1992. Treasurer Costello was quick to credit the baby bonus: The daily newspaper *The Australian* quoted Costello as "delighted that at least some families have been taking up the challenge."

ANU's McDonald says that although it is too early to carry out "rigorous research" on the reasons for the increase, most of the additional births are to women in the middle to late part of their childbearing years. This suggests that the message may have been heard: "If you want to have children, it is risky to delay too long," McDonald says. And although McDonald concedes that "most of the 261,000 women who gave birth in 2005 would have had the baby without" the baby bonus, the extra money "can make a difference" to middle income families who make "close calculations" about the impact of parenting. McDonald estimates that Australia's TFR for 2005, when published in November, should rise from about 1.76 to 1.82.

But Robert Birrell, director of the Center for Population and Urban Research at Monash University in Clayton, Australia, says that a number of other factors may have weighed much more heavily, especially "the impact of the current economic boom in Australia, which has seen an increase in the rate of employment for men and particularly women in recent years." Santow agrees: "I would not leap immediately to the conclusion that Peter Costello should be given the credit."

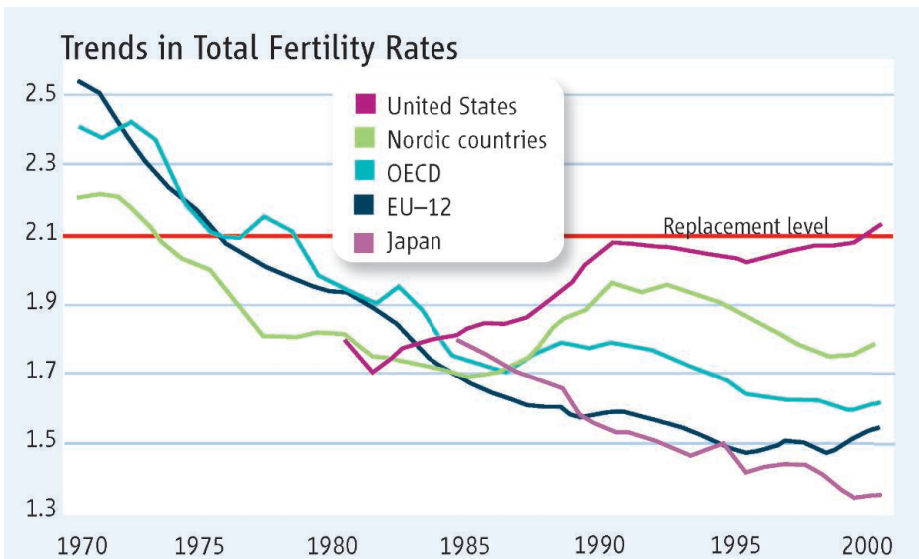
### Low-fertility spiral

The uncertain response to incentives suggests to some demographers that governments need to do even more to make child rearing attractive. "Many things that we've tried aren't big enough," says Bongaarts. "To move behavior, you need real incentives; you need thousands of dollars. … You have to pull all the levers you have, and maybe you will get halfway there." But pulling those levers might end up being too costly, Trussell says. "Policies that would work would be so expensive that they will never be implemented."

And some researchers have begun to think that it might actually be too late to reverse the trend in countries with the lowest fertility levels. At several recent population meetings, for example, McDonald has warned that once a nation's TFR falls below 1.5, a downward demographic spiral sets in that makes it much more difficult to recover. "This is the safety zone," McDonald says. "Countries should try hard to avoid falling below it."

A team led by Wolfgang Lutz, a demographer at the International Institute for Applied Systems Analysis (IIASA) in Laxenburg, Austria, has taken McDonald's observation further and argued that countries with a TFR of 1.5 or lower may have crossed into permanent negative population growth. Lutz calls this hypothesis, which he presented most recently at this spring's annual meeting of the Population Association of America in Los Angeles, California, the "low-fertility trap." Lutz and other colleagues at IIASA and the Vienna Institute of Demography argue that the new social norms created by low fertility rates create a self-reinforcing negative feedback loop. It is locked in place by a reduction in ideal family size, aging of the population, and other effects on the labor market that make having fewer and fewer children inevitable. As evidence, Lutz and his colleagues cite data from the Eurobarometer survey showing that in Germany and Austria—nations with TFRs of 1.39 and 1.36, respectively—young adults now consider their ideal family sizes to be as low as 1.7 children on average.

"Germany is the extreme example of this phenomenon, with around 30% of young people not intending to have children," says McDonald.




SOURCE: OECD

On the other hand, McDonald does not agree that there is no turning back for countries whose TFRs fall this low: "This does make Germany a tougher nut to crack, but I would never declare the game as over."

Yet Reher sees little reason for optimism. "When fertility is drastically below replacement, it doesn't go up, no matter how many policies and how much money is thrown at it," he says. "We are in the midst of a cascading fertility decline. Even a TFR of 1.7 is not safe; it is a disaster if you look a couple of generations down the line."

Indeed, Reher, at the July 2005 annual meeting of the International Union for the Scientific Study of Population in Tours, France, presented a paper suggesting an even more dismal picture. Reher argued that low fertility rates were now entrenched in the social structure of developed countries and a growing number of developing countries as well. Although the momentum of past high fertility rates would continue to fuel an increase in the entire world's population for some decades to come, this would eventually stop. Rather, Reher maintained, much of the world is now on the cusp of a prolonged period of population decline. The resulting population aging would lead to labor shortages even in developing countries. The result could be an economic disaster, Reher warned. "Urban areas in regions like Europe could well be filled with empty buildings and crumbling infrastructures as population and tax revenues decline," he prognosticated, adding that "it is not difficult to imagine enclaves of rich, fiercely guarded pockets of well-being surrounded by large areas which look more like what we might see in some science-fiction movies."

Most population researchers agree that there is plenty to worry about in current worldwide demographic trends. Yet few are ready to accept the direst parts of Reher's doomsday scenario—at least not yet. "I wouldn't be surprised" if population shrinkage "happens in a lot of places in the world," says Gauthier, although she adds that "it is much harder to believe in Africa," where the population is expected to at least double by 2050. And Santow comments that although Reher's predictions "may well be sensible," she sees "nothing terrifying about a drop in the size of Europe's population. Any decline will take time, and economies will adjust. Governments should not expend energy to maintain the status quo. Governments should plan for the future, not try to reintroduce the past."

**–MICHAEL BALTER**


Published by AAAS

PERSPECTIVE

# Some Frontiers in Social Science

William P. Butz* and Barbara Boyle Torrey

The fundamental challenge in the social sciences is moving from complicated correlations to useful prediction. Progress usually reflects an interplay between theory, data, and tools. Six areas of innovation, principally data and tools, are now pushing at the frontiers of these sciences: longitudinal data, laboratory experimentation, improved statistical methods, geographic information tools, biosocial science, and international replication. These innovations are gaining power as they cross disciplinary boundaries, helping to attribute causality to observed relationships, to understand their nature, and thereby to improve the accuracy and usefulness of predictions.

The fundamental challenge in the social sciences is moving from complicated correlations to useful prediction (*1*). The variance in human behavior and social adaptations is so large that it becomes difficult to distinguish signals from noise in data, even where theory supplies clear guidance. However, new data, methodologies, and tools from both inside and outside the social sciences are demonstrating real promise in advancing these sciences from descriptive to predictive ones (*2*).

### Longitudinal Surveys

Longitudinal surveys, which collect information about the same persons over many years, have given the social sciences their Hubble telescope. Both allow the observing researcher to look back in time and record the antecedents of current events and transitions. These longitudinal data sets, beginning 40 years ago with the Panel Study of Income Dynamics (*3*), document the importance of accumulated life experience in causing transitions from health to infirmity; from work to unemployment or retirement; and across states of marriage, family structure, and wealth.

Similar longitudinal surveys in different countries are beginning to reveal the institutional and cultural variations in these life cycle dynamics (*4*). A new investment in international longitudinal surveys on aging may eventually unravel the changing dynamics at the end of the life cycle across time and countries (*5*). One recent comparison showed that health improved with socio-economic status in both England and the United States. It also showed, however, that English health status is better than in the United States at each socio-economic level (*6*). The challenge now is to use the accumulated longitudinal data on individuals to explain these differences in health.

### Laboratory Experimentation

The new field of behavioral social science integrates insights from across the social and behavioral sciences with data from the neurosciences to better understand economic, social, and political outcomes (*7*). An important tool of behavioral social science is controlled laboratory experimentation, usually conducted at computers where subjects respond to specific instructions and online stimuli. Game theory, risk and decision science, and experimental social psychology have provided most of the theory and methods for this new approach (*8*). The objects of inquiry span all the social sciences, including the origins and impacts of ethnic conflict (*9*), group and team behavior in organizations (*10*), and the nature and consequences of trust and reciprocity in interpersonal and international relations. By manipulating participants' instructions, incentives, constraints, and rewards in an experimental setting, investigators can fine-tune the parameters of real-world applications. Such applications are by now diverse and their use has spread around the globe: broadband spectrum auctions (*11*), tradable permits to pollute (*12*), airport landing rights algorithms, voucher plans for schools, and income-contingent loans for higher education. One current challenge among many is designing mechanisms to induce development and distribution of vaccines for neglected diseases that affect poor countries (*13*).

More generally, research is now testing whether behavior in a controlled game indeed predicts future real-world behavior of the individuals who played the game (*14*), whether such behavior corresponds to evidence from independent sample surveys, and whether field experiments outside the laboratory support the findings (*15*). As Fig. 1 illustrates, better to experiment with different settings of the "life machine" in the laboratory before imposing them through organizations and governments on the real lives of people.

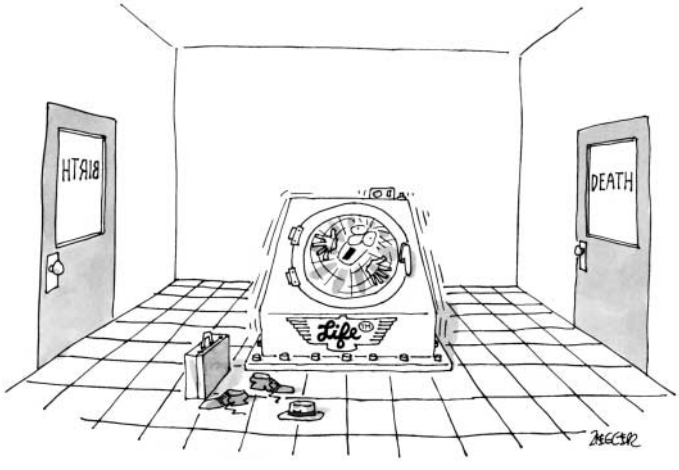



**Fig. 1.** (© 2000) Jack Zielgler/The New Yorker Collection/Cartoonbank.com

### Improved Statistical Methods

The Internet provides ready and inexpensive access around the world to hundreds of thousands of potential survey respondents, many more than even the largest conventional survey can include (*16*). Such respondents could also be experimental subjects, reacting to the experimental design in real time and in unprecedented numbers. These respondents could be preselected as a controlled sample, but greater potential lies in granting easy access to anyone who wants to participate, resulting in huge "convenience samples" that do not represent any known larger population. The challenge, to draw proper inferences to a universe larger and more interesting than just the volunteer survey respondents, has been well known for decades in the context of program- and clinic-based data. What is different now is the suddenly large payoff from solving this challenge of respondent selection bias (*17*). A promising approach is to reweight the data based on the known distribution of some key variables that are thought to capture the difference between Web survey respondents and the target population. The idea is that the weights used to correct for the discrepancy to the known distribution of the key variables will also remove or reduce bias for other variables in the survey (*18*, *19*). Another approach is to test theories on multiple independent Web samples (*20*).

There is a debilitating tradeoff between the power of richly detailed data on individuals (and firms and other organizations) to test important hypotheses, on the one hand, and the possibility that such detailed information will threaten the privacy and confidentiality of the persons and organizations described in the data, on the other hand. The masking of subject iden-

Population Reference Bureau, 1875 Connecticut Avenue, Washington, DC 20009, USA.

*To whom correspondence should be addressed. E-mail: wbutz@prb.org

tity arises in many fields of science but is particularly vexing in the social sciences where available masking techniques necessarily reduce the amount of information in the data. Exact-matched data files from multiple surveys and administrative record sources are particularly valuable analytically, but also particularly at risk for disclosure. Consider, for example, matched files that contain demographic and attitudinal information from persons in surveys, economic information on these same persons from government records, and medical information from providers. When longitudinal data are so matched, the analytical payoff can be unprecedented at relatively little cost in time and money.

Some researchers now have increased access to such data through cumbersome administrative and legal arrangements, such as traveling to special protected facilities or submitting statistical software to be run in a protected facility (*21*, *22*). These challenges arise particularly but not exclusively with data files collected by the federal government. Statistical methods of masking subject identities while retaining sufficient information for analysis are under development (*23*). Breakthroughs here will have substantial scientific payoff.

### Geographic Information Tools

The context of human social interaction has always been critical for understanding social outcomes. The integration of Geographic Information Science (GIS) and Geographic Positioning Systems (GPS) with social science is beginning to socialize the pixel by providing a geographic context for social behavior (*24*). These new tools provide the ability to analyze social behavior across time and geographic scales, although their adoption by social scientists has yet to approach their potential. The integration of geography and macroeconomics will allow researchers to estimate how much of the regional differences in per capita output and output per geographic area is due to country-specific factors, such as institutional differences, and how much to other factors, such as geography (*25*). This integration may eventually disentangle the explanations for why some countries and regions within countries lag economically. In addition, the global distribution of population by elevation will provide future estimates of coastal and biological hazards as well as insights into human settlement patterns (*26*). Spatial integration of population change and consumption patterns may also become increasingly important in understanding environmental and climate changes (*27*).

### Biosocial Science

No field outside the social sciences is having as much of an impact within the field as biology. This new biosocial science may well alter understandings about sexual orientation, criminal responsibility, prospects for marriage as a social institution, and even the nature of moral obligation.

Mapping of the human genome has contributed fundamentally to understanding how the human body has evolved, how it operates, and how it malfunctions. Genomics has also provided the tools to understand important social science phenomena such as prehistoric human migrations (*28*). The International Hap Map project will go much further in exploring the nature of genetic variation among human beings. This project is challenging the racial and ethnic constructs that much of social science has used from its beginnings (*29*). Although determinants of human behavior are both genetic and environmental, biosocial research may also alter our understanding of their relative impacts.

Neurosocial science uses neuroimaging to study the origins and nature of economic, social, and political behavior, examining such phenomena as decision-making, empathy, and time preference. The work is mainly descriptive so far but is beginning to suggest fundamental distinctions between behaviors that have been assumed to have identical causes and characteristics. For example, two separate neural systems are involved in decisions between monetary award options, depending on time to delivery (*30*). There is also evidence for a neurobiological link between the experienced displeasure of dread and subsequent decisions about unpleasant outcomes (*31*).

Social scientists are also collaborating with animal behaviorists to test the species generality of hypotheses originally generated to explain human behavior. The observations that animals can count, plan ahead, and make decisions under uncertainty, and that most other primate females also live longer than males, call into question some long-standing beliefs about "human" behavior (*32*).

Field surveys of economic, social, and health phenomena are increasingly incorporating biomarkers collected from the survey respondents, in order to study biosocial interactions in fertility, work, health, and aging (*33*). Indeed, biomarkers on social surveys may well reveal more about subjects' predispositions and their ancestry than do their verbal responses on which social scientists have historically depended.

Over the past two decades, the theory of evolution has influenced parts of economics and psychology, and to a lesser extent sociology, anthropology, and political science (*34*, *35*). Where do our behaviors and their reflections in our organizations (including the family) come from? What constraints on our "freedom" do our hard-wired tendencies impose? Are there cross-cultural differences? What refutable predictions does an evolutionary approach imply that can be tested against other theories? What are the practical and policy implications? These questions, largely unanswered, are fundamentally important.

### International Replications

International replication of national findings is critical for distinguishing between local and universal phenomena. The recent development of internationally harmonized cross-sectional data sets has already demonstrated their potential to generate hypotheses for future research. The Demographic and Health Surveys have documented fertility declines in developing countries, but have raised major questions about why the patterns of decline vary so much (*36*). The Luxembourg Income Study has shown the income vulnerability of single-parent families in developed countries but has raised questions about why some countries are so much more efficient at compensating for the vulnerability than others (*37*). The World Value Survey has documented the trends toward secularism in developed countries over the past two decades, which has raised questions about why the United States is an exception (*38*). Some future breakthroughs in understanding social dynamics will begin with questions raised by these kinds of cross-sectional comparisons.

### Conclusions

Progress in the social sciences usually reflects an interplay between ideas, data, and tools. In our view, this interplay at the frontiers has in recent decades been driven more by new data, tools, and methods and less by theoretical advances and novel hypotheses, although in some areas these have been plentiful. Many of these innovations have gained power as they crossed disciplinary boundaries, some from outside the social sciences entirely. Those discussed here have more potential than has been realized so far to attribute causality to observed relationships, to understand their nature, and thereby to improve the accuracy and usefulness of predictions.

#### References and Notes


1. J. H. Marberger, keynote address presented at the Conference on Science and Technology Policy, Atlanta, GA, 18 to 20 May 2006.
2. For earlier reviews of social science frontiers, see (*39*, *40*).
3. Panel Study of Income Dynamics (psidonline.isr.umich.edu/guide/overview.html).
4. B. Bradury, S. P. Jenkins, J. Micklewright, *The Dynamics of Child Poverty in Industrialized Countries* (Cambridge Univ. Press, New York, 2001).
5. R. Suzman, *Popul. Dev. Rev. Suppl.* **30**, 239 (2004).
6. J. Banks, M. Marmot, Z. Oldfield, J. P. Smith, *JAMA* **295**, 2037 (2006).
7. C. F. Camerer, G. Loewenstein, R. Rabin, Eds., *Advances in Behavioral Economics* (Princeton Univ. Press, Princeton, NJ, 2003).
8. D. McFadden, *Am. Econ. Rev.* **96**, 5 (2006).
9. J. D. Fearon, D. D. Laitin, *Am. Polit. Sci. Rev.* **90**, 715 (1996).
10. R. A. Weber, *Am. Econ. Rev.* **96**, 114 (2006).
11. The number and variety of auction programs are rapidly growing (*41*).
12. So-called designer markets now operate for many goods or "bads" (such as pollution) in many countries where, for whatever reason, markets did not naturally develop. Designer markets typically require considerable laboratory experimentation to get the incentives right for all actors—buyers, sellers, and the public interest—before they are put into practice.
13. E. Berndt *et al.*, "Advanced purchase commitments for a malaria vaccine: Estimating costs and effectiveness" (National Bureau of Economic Research, NBER Working Paper 11288, Cambridge, MA, 2005).

14. D. S. Karlan, *Am. Econ. Rev.* **95**, 1688 (2005).
15. G. W. Harrison, J. A. List, *J. Econ. Lit.* **42**, 1009 (2004).
16. D. A. Dillman, *Mail and Internet Surveys: The Tailored Design Method* (Wiley, New York, 2000).
17. Beyond this statistical challenge, the Web is changing how social science, along with all of science, is conducted. For example, massive records of the Web transactions themselves are data for analysis that uses complexity theory and network theory to understand social and economic networks (*42*, *43*).
18. J. Witte in *Society Online: The Internet in Context*, P. N. Howard, S. Jones, Eds. (Russell Sage Foundation, New York, 2004), p. xv.
19. M. Schonlau, A. Van Soest, A. Kapteyn, M. Couper, J. Winter, in preparation.
20. W. S. Bainbridge in *Computing in the Social Sciences*, O. V. Burton, Ed. (University of Illinois Press, Urbana, IL, 2002), pp. 51–56.
21. U.S. Census Bureau Center for Economic Studies (www.ces.census.gov).
22. Procedures and Costs for Use of the Research Data Center (www.cdc.gov/nchs/r&d/rdc.htm).
23. J. M. Abowd, J. I. Lane, *Tech. Pap. No. TP-2003-10* (U.S. Census Bureau, Washington, DC, 2003).
24. D. Livermore, E. F. Moran, R. R. Rindfuss, P. C. Stern, Eds., *People and Pixels* (National Academy Press, Washington, DC, 1998).
25. W. D. Nordhaus, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 3510 (2006).
26. J. E. Cohen, C. Small, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 14009 (1998).
27. B. C. O'Neill, F. L. MacKellar, W. Lutz, *Population and Climate Change* (Cambridge Univ. Press, Cambridge, 2001), pp. 114–117.
28. L. Jin *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 3796 (1999).
29. V. L. Bonham, E. Warshauer-Baker, F. S. Collins, *Am. Psychol.* **60**, 9 (2005).
30. S. M. McClure, D. I. Laibson, G. Loewenstein, J. D. Cohen, *Science* **306**, 503 (2004).
31. G. S. Berns *et al.*, *Science* **312**, 754 (2006).
32. J. H. Kagel, R. C. Battalio, L. Green, *Economic Choice Theory: An Experimental Analysis of Animal Behavior* (Cambridge Univ. Press, Cambridge, 1995).
33. Chicago Core on Biomarkers in Population-Based Aging Research (http://biomarkers.uchicago.edu/studiescollectingbiomarkers.htm) summarizes studies collecting biomarkers in population settings.
34. J. Henrich, *Science* **312**, 60 (2006).
35. D. C. Dennett, *Breaking the Spell: Religion as a Natural Phenomenon* (Viking Press, New York, 2006).
36. Demographic and Health Surveys (www.measuredhs.com).
37. Luxembourg Income Study (www.lisproject.org).
38. World Values Survey (www.worldvaluessurvey.org).
39. K. W. Deutsch, J. Platt, D. Senghaas, *Science* **171**, 450 (1971).
40. P. M. Smith, B. B. Torrey, *Science* **271**, 611 (1996).
41. P. Milgrom, *Putting Auction Theory to Work* (Cambridge Univ. Press, Cambridge, 2004).
42. M. E. J. Newman, A. L. Barabasdi, D. J. Watts, Eds., *The Structure and Dynamics of Complex Networks* (Princeton Univ. Press, Princeton, NJ, 2003).
43. D. J. Watts, *Six Degrees: The Science of a Connected Age* (Norton, New York, 2003).


44. W.P.B. is a member of the Board of Reviewing Editors of *Science* and former Director of the Division of Social, Economic, and Behavioral Research at the NSF. B.B.T. is a Fellow of the AAAS and former Executive Director of the Commission on Behavioral and Social Sciences and Education at the National Research Council.

10.1126/science.1130121

PERSPECTIVE

# Skill Formation and the Economics of Investing in Disadvantaged Children

**James J. Heckman**

**This paper summarizes evidence on the effects of early environments on child, adolescent, and adult achievement. Life cycle skill formation is a dynamic process in which early inputs strongly affect the productivity of later inputs.**

Four core concepts important to devising sound social policy toward early childhood have emerged from decades of independent research in economics, neuroscience, and developmental psychology (*1*). First, the architecture of the brain and the process of skill formation are influenced by an interaction between genetics and individual experience. Second, the mastery of skills that are essential for economic success and the development of their underlying neural pathways follow hierarchical rules. Later attainments build on foundations that are laid down earlier. Third, cognitive, linguistic, social, and emotional competencies are interdependent; all are shaped powerfully by the experiences of the developing child; and all contribute to success in the society at large. Fourth, although adaptation continues throughout life, human abilities are formed in a predictable sequence of sensitive periods, during which the development of specific neural circuits and the behaviors they mediate are most plastic and therefore optimally receptive to environmental influences.

A landmark study concluded that "virtually every aspect of early human development, from the brain's evolving circuitry to the child's capacity for empathy, is affected by the environments and experiences that are encountered in a cumulative fashion, beginning in the prenatal period and extending throughout the early childhood years" (*2*). This principle stems from two characteristics that are intrinsic to the nature of learning: (i) early learning confers value on acquired skills, which leads to self-reinforcing motivation to learn more, and (ii) early mastery of a range of cognitive, social, and emotional competencies makes learning at later ages more efficient and therefore easier and more likely to continue.

Early family environments are major predictors of cognitive and noncognitive abilities. Research has documented the early (by ages 4 to 6) emergence and persistence of gaps in cognitive and noncognitive skills (*3*, *4*). Environments that do not stimulate the young and fail to cultivate these skills at early ages place children at an early disadvantage. Disadvantage arises more from lack of cognitive and noncognitive stimulation given to young children than simply from the lack of financial resources.

This is a source of concern because family environments have deteriorated. More U.S. children are born to teenage mothers or are living in single parent homes compared with 40 years ago (*5*). Disadvantage is associated with poor parenting practices and lack of positive cognitive and noncognitive stimulation. A child who falls behind may never catch up. The track records for criminal rehabilitation, adult literacy, and public job training programs for disadvantaged young adults are remarkably poor (*3*). Disadvantaged early en-

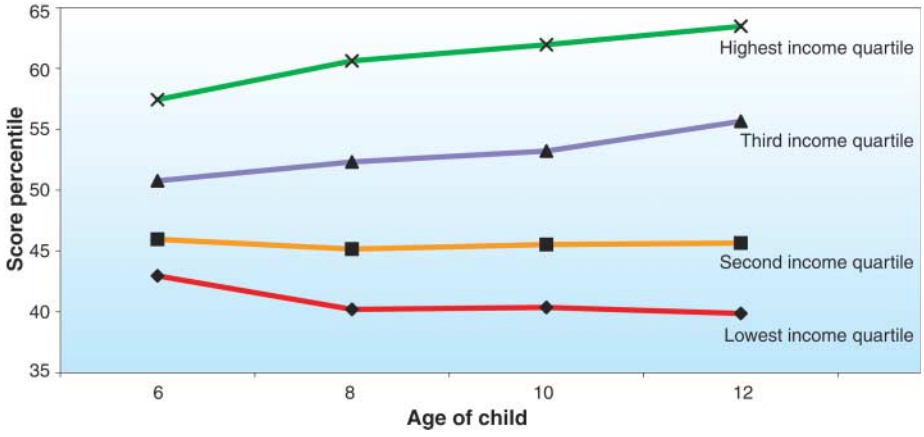



**Fig. 1.** Average percentile rank on Peabody Individual Achievement Test–Math score by age and income quartile. Income quartiles are computed from average family income between the ages of 6 and 10. Adapted from (*3*) with permission from MIT Press.

Department of Economics, University of Chicago, Chicago, IL 60637, USA. Department of Economics, University College Dublin, Dublin 4, Ireland. E-mail: jjh@uchicago.edu

vironments are powerful predictors of adult failure on a number of social and economic measures.

Many major economic and social problems can be traced to low levels of skill and ability in the population. The U.S. will add many fewer college graduates to its workforce in the next 20 years than it did in the past 20 years (*6*, *7*). The high school dropout rate, properly measured with inclusion of individuals who have received general educational development (GED) degrees, is increasing at a time when the economic return of schooling has increased (*8*). It is not solely a phenomenon of unskilled immigrants. Over 20% of the U.S. workforce is functionally illiterate, compared with about 10% in Germany and Sweden (*9*). Violent crime and property crime levels remain high, despite large declines in recent years. It is estimated that the net cost of crime in American society is $1.3 trillion per year, with a per capita cost of $4818 per year (*10*). Recent research documents the importance of deficits in cognitive and noncognitive skills in explaining these and other social pathologies (*11*).

### Noncognitive Skills and Examples of Successful Early Interventions

Cognitive skills are important, but noncognitive skills such as motivation, perseverance, and tenacity are also important for success in life. Much public policy, such as the No Child Left Behind Act, focuses on cognitive test score outcomes to measure the success of interventions in spite of the evidence on the importance of noncognitive skills in social success. Head Start was deemed a failure in the 1960s because it did not raise the intelligence quotients (IQs) of its participants (*12*). Such judgments are common but miss the larger picture. Consider the Perry Preschool Program (*13*), a 2-year experimental intervention for disadvantaged African-American children initially ages 3 to 4 that involved morning programs at school and afternoon visits by the teacher to the child's home. The Perry intervention group had IQ scores no higher than the control group by age 10. Yet, the Perry treatment children had higher achievement test scores than the control children because they were more motivated to learn. In followups to age 40, the treated group had higher rates of high school graduation, higher salaries, higher percentages of home ownership, lower rates of receipt of welfare assistance as adults, fewer out-of-wedlock births, and fewer arrests than the controls (*13*). The economic benefits of the Perry Program are substantial (Table 1). Rates of return are 15 to 17% (*14*). (The rate of return is the increment in earnings and other outcomes, suitably valued, per year for each dollar invested in the child). The benefit-cost ratio (the ratio of the aggregate program benefits over the life of the child to the input costs) is over eight to one.

Perry intervened relatively late. The Abecedarian program, also targeted toward disadvantaged children, started when participants were 4 months of age. Children in the treatment group received child care for 6 to 8 hours per day, 5 days per week, through kindergarten entry; nutritional supplements, social work services, and medical care were provided to control group families. The program was found to permanently raise the IQ and the noncognitive skills of the treatment group over the control group. However, the Abecedarian program was intensive, and it is not known whether it is the age of intervention or its intensity that contributed to its success in raising IQ (*15*–*17*).

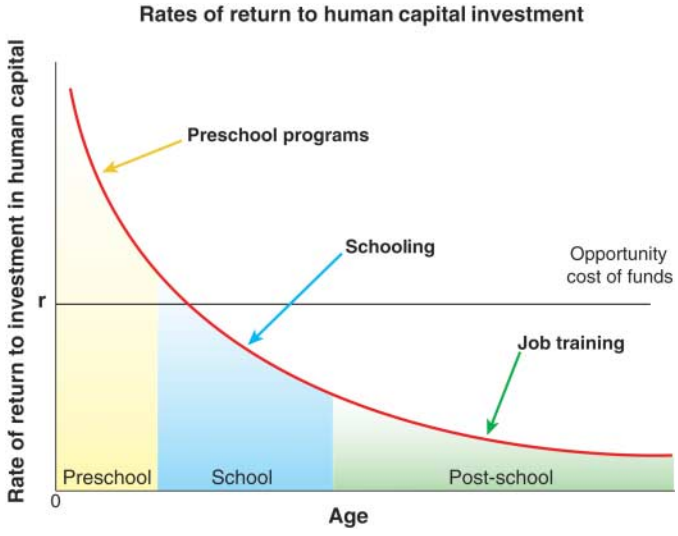



**Fig. 2.** Rates of return to human capital investment in disadvantaged children. The declining figure plots the payout per year per dollar invested in human capital programs at different stages of the life cycle for the marginal participant at current levels of spending. The opportunity cost of funds (*r*) is the payout per year if the dollar is invested in financial assets (e.g., passbook savings) instead. An optimal investment program from the point of view of economic efficiency equates returns across all stages of the life cycle to the opportunity cost. The figure shows that, at current levels of funding, we overinvest in most schooling and post-schooling programs and underinvest in preschool programs for disadvantaged persons. Adapted from (*3*) with permission from MIT Press.

Reynolds *et al.* present a comprehensive review of early childhood programs directed toward disadvantaged children and their impact (*18*). Similar returns are obtained for other early intervention programs (*19*, *20*), although more speculation is involved in these calculations because the program participants are in the early stages of their life cycles and do not have long earnings histories.

### Schools and Skill Gaps

Many societies look to the schools to reduce skills gaps across socioeconomic groups. Because of the dynamics of human skill formation, the abilities and motivations that children bring to school play a far greater role in promoting their performance in school than do the traditional inputs that receive so much attention in public policy debates. The Coleman Report (*21*) as well as recent work (*22*, *23*) show that families and not schools are the major sources of inequality in student performance. By the third grade, gaps in test scores across socioeconomic groups are stable by age, suggesting that later schooling and variations in schooling quality have little effect in reducing or widening the gaps that appear before students enter school (*4*, *24*). Figure 1 plots gaps in math test scores by age across family income levels. The majority of the gap at age 12 appears at the age of school enrollment. Carneiro and Heckman performed a cost-benefit analysis of classroom size reduction on adult earnings (*3*). Although smaller classes raise the adult earnings of students, the earnings gains received by students do not offset the costs of hiring additional teachers. The student-teacher achievement ratio (STAR) randomized trial of classroom size in Tennessee shows some effect of reduced classroom size on test scores and adult performance, but most of the effect occurs in the earliest grades (*25*, *26*). Schools and school quality at current levels of funding contribute little to the emergence of test score gaps among children or to the development of the gaps.

### Second Chance Programs

America is a second chance society. Our educational policy is based on a fundamental optimism about the possibility of human change. The dynamics of human skill formation reveal that later compensation for deficient early family environments is very costly (*4*). If society waits too long to compensate, it is economically inefficient to invest in the skills of the disadvantaged. A serious trade-off exists between equity and efficiency for adolescent and young adult skill policies. There is no such trade-off for policies targeted toward disadvantaged young children (*28*).

The findings of a large literature are captured in Fig. 2. This figure plots the rate of return, which is the dollar flow from a unit of investment at each age for a marginal investment in a disadvantaged young child at current levels of expenditure. The economic return from early interventions is high, and the return from later interventions is lower. Remedial programs in the adolescent and young adult years are much more costly in producing the same level of skill attainment in adulthood. Most are economically inefficient. This is reflected in Fig. 2 by the fact that a segment of the curve lies below the opportunity cost of funds (the horizon-

tal line fixed at $r$). The opportunity cost is the return from funds if they were invested for purposes unrelated to disadvantaged children.

### Conclusions

Investing in disadvantaged young children is a rare public policy initiative that promotes fairness and social justice and at the same time promotes productivity in the economy and in society at large. Early interventions targeted toward disadvantaged children have much higher returns than later interventions such as reduced pupil-teacher ratios, public job training, convict rehabilitation programs, tuition subsidies, or expenditure on police. At current levels of resources, society overinvests in remedial skill investments at later ages and underinvests in the early years.

Although investments in older disadvantaged individuals realize relatively less return overall, such investments are still clearly beneficial. Indeed, the advantages gained from effective early interventions are sustained best when they are followed by continued high-quality learning experiences. The technology of skill formation shows that the returns on school investment and postschool investment are higher for persons with higher ability, where ability is formed in the early years. Stated simply, early investments must be followed by later investments if maximum value is to be realized.

**Table 1.** Economic benefits and costs of the Perry Preschool Program (*27*). All values are discounted at 3% and are in 2004 dollars. Earnings, Welfare, and Crime refer to monetized value of adult outcomes (higher earnings, savings in welfare, and reduced costs of crime). K–12 refers to the savings in remedial schooling. College/adult refers to tuition costs.

| | Perry Preschool |
|---|---|
| Child care | $986 |
| Earnings | $40,537 |
| K–12 | $9184 |
| College/adult | $–782 |
| Crime | $94,065 |
| Welfare | $355 |
| Abuse/neglect | $0 |
| Total benefits | $144,345 |
| Total costs | $16,514 |
| Net present value | $127,831 |
| Benefits-to-costs ratio | 8.74 |

### References and Notes


1. E. I. Knudsen, J. J. Heckman, J. Cameron, J. P. Shonkoff, *Proc. Natl. Acad. Sci. U.S.A.*, in press.
2. J. P. Shonkoff, D. Phillips, *From Neurons to Neighborhoods: The Science of Early Child Development* (National Academies Press, Washington, DC, 2000).
3. P. Carneiro, J. J. Heckman, in *Inequality in America: What Role for Human Capital Policies?* J. J. Heckman, A. B. Krueger, B. Friedman, Eds. (MIT Press, Cambridge, MA, 2003), ch. 2, pp. 77–237.
4. F. Cunha, J. J. Heckman, L. J. Lochner, D. V. Masterov, in *Handbook of the Economics of Education*, E. A. Hanushek, F. Welch, Eds. (North Holland, Amsterdam, in press).
5. J. J. Heckman, D. V. Masterov, "The productivity argument for investing in young children," (Working Paper No. 5, Committee on Economic Development, Washington, DC, 2004).
6. J. B. Delong, L. Katz, C. Goldin, in *Agenda for the Nation*, H. Aaron, J. Lindsay, P. Nivola, Eds. (Brookings Institution Press, Washington, DC, 2003), pp. 17–60.
7. D. T. Ellwood, in *The Roaring Nineties: Can Full Employment Be Sustained?* A. Krueger, R. Solow, Eds. (Russell Sage Foundation, New York, 2001), pp. 421–489.
8. J. J. Heckman, P. LaFontaine, *J. Lab. Econ.*, in press.
9. *International Adult Literacy Survey, 2002: User's Guide*, Statistics Canada, Special Surveys Divison, National Literacy Secretariat, and Human Resources Development Canada (Statistics Canada, Ottawa, Ontario, 2002).
10. D. A. Anderson, *J. Law Econ.* **42**, 611 (1999).
11. J. J. Heckman, J. Stixrud, S. Urzua, *J. Lab. Econ.*, in press.
12. Westinghouse Learning Corporation and Ohio University, *The Impact of Head Start: An Evaluation of the Effects of Head Start on Children's Cognitive and Affective Development*, vols. 1 and 2 (Report to the Office of Economic Opportunity, Athens, OH, 1969).
13. L. J. Schweinhart *et al.*, *Lifetime Effects: The High/Scope Perry Preschool Study Through Age 40* (High/Scope, Ypsilanti, MI, 2005).
14. A. Rolnick, R. Grunewald, "Early childhood development: Economic development with a high public return" (Tech. rep., Federal Reserve Bank of Minneapolis, Minneapolis, MN, 2003).
15. C. T. Ramey, S. L. Ramey, *Am. Psychol.* **53**, 109 (1998).
16. C. T. Ramey, S. L. Ramey, *Prev. Med.* **27**, 224 (1998).
17. C. T. Ramey *et al.*, *Appl. Dev. Sci.* **4**, 2 (2000).
18. A. J. Reynolds, M. C. Wang, H. J. Walberg, *Early Childhood Programs for a New Century* (Child Welfare League of America Press, Washington, DC, 2003).
19. L. A. Karoly *et al.*, *Investing in Our Children: What We Know and Don't Know About the Costs and Benefits of Early Childhood Interventions* (RAND, Santa Monica, CA, 1998).
20. L. N. Masse, W. S. Barnett, *A Benefit Cost Analysis of the Abecedarian Early Childhood Intervention* (Rutgers University, National Institute for Early Education Research, New Brunswick, NJ, 2002).
21. J. S. Coleman, *Equality of Educational Opportunity* (U.S. Deparment of Health, Education, and Welfare, Office of Education, Washington, DC, 1966).
22. S. W. Raudenbush, "Schooling, statistics and poverty: Measuring school improvement and improving schools" Inaugural Lecture, Division of Social Sciences, University of Chicago, Chicago, IL, 22 February 2006.
23. J. J. Heckman, M. I. Larenas, S. Urzua, unpublished data.
24. D. A. Neal, in *Handbook of Economics of Education*, E. Hanushek, F. Welch, Eds. (Elsevier, Amsterdam, in press).
25. B. Krueger, D. M. Whitmore, *Econ. J.* **111**, 1 (2001).
26. B. Krueger, D. M. Whitmore, in *Bridging the Achievement Gap*, J. E. Chubb, T. Loveless, Eds. (Brookings Institution Press, Washington, DC, 2002).
27. W. S. Barnett, Benefit-Cost Analysis of Preschool Education, 2004, (http://nieer.org/resources/files/BarnettBenefits.ppt).
28. F. Cunha, J. J. Heckman, *J. Hum. Resour.*, in press.


29. This paper was generously supported by NSF (grant nos. SES-0241858 and SES-0099195), National Institute of Child Health and Human Development (NIH grant no. R01HD043411), funding from the Committee for Economic Development, with a grant from the Pew Charitable Trusts and from the Partnership for America's Economic Success. This research was also supported by the Children's Initiative project at the Pritzker Family Foundation and a grant from the Report to the Nation of America's Promise. The views expressed in this paper are those of the author and not necessarily those of the sponsoring organizations. See our Web site (http://jenni.uchicago.edu/econ_neurosci) for more information.

10.1126/science.1128898


PERSPECTIVE

# Studying Adolescence

**Linda M. Richter**

Young people in their teens constitute the largest age group in the world, in a special stage recognized across the globe as the link in the life cycle between childhood and adulthood. Longitudinal studies in both developed and developing countries and better measurements of adolescent behavior are producing new insights. The physical and psychosocial changes that occur during puberty make manifest generational and early-childhood risks to development, in the form of individual differences in aspects such as growth, educational attainment, self-esteem, peer influences, and closeness to family. They also anticipate threats to adult health and well-being. Multidisciplinary approaches, especially links between the biological and the social sciences, as well as studies of socioeconomic and cultural diversity and determinants of positive outcomes, are needed to advance knowledge about this stage of development.

Young people aged 10 to 19 currently constitute a demographic bulge. They are the largest age group in the world, making up close to 20% of the 6.5 billion world population estimated in 2005 (*1*), 85% of whom live in developing countries and account for about one-third of those countries' national populations. Adolescence has also been described as "demographically dense": a period in life during which a large percentage of people experience a large percentage of key life-course events (*2*). These include leaving or completing school, bearing a child, and becoming economically productive. They also include experiences, more common in this age group than in others, that are capable of substantially altering life trajectories: nonconsensual sex, alcohol and drug abuse, self-harm and interpersonal violence, and getting into trouble with the law. Diet and activity patterns, friendships, educational achievement, and civic involvement all affect current health,

Child, Youth, Family, and Social Development, Human Sciences Research Council, Private Bag X07, Dalbridge 4014, South Africa, and University of KwaZulu-Natal, South Africa. E-mail: lrichter@hsrc.ac.za

schooling, and family life, but they also have long-term effects on well-being in adulthood and even on future generations.

Adolescence was long thought to be an American cultural invention, a by-product of industrialization; a personally and socially problematic period created by the dependence of young people on their parents after leaving school and waiting to find work (*3*). However, investigations of hundreds of societies confirm that adolescence is a universally recognized life stage, starting around or just after puberty, although with different markers, behavioral manifestations, and social attributions (*4*–*7*). In most societies, the onset of adolescence is celebrated through rituals associated with prospective adult roles, such as reproduction, responsibility, and work; or with religious ceremonies, often differentiated by gender. In highly industrialized societies, the rites of passage are less public and more variable, but the period between childhood and adulthood may be bridged by changes in schooling, changes in family rules about autonomy, "first-time" experiences such as drinking alcohol, or inauguration into a group (*8*, *9*).

After the publication in 1904 of a widely influential book by Hall (*10*), there was a pervasively held opinion that the "raging hormones" of puberty inevitably led to rebellion, conflict with parents and other authorities, etc. This simplistic idea has been shown to be false (*9*). Most adolescents don't go through a period of "sturm und drang." Rather, the period of transition to adulthood—largely a socially designated period, with an onset that is generally precipitated by apparent physical changes associated with puberty—involves multiple interactions between biology and culture (or the set of social institutions and relationships prevalent at the time). Unidirectional models that assume that hormones cause behavior (for example, that testosterone causes aggression) or that behaviors cause hormone change (for example, that stress increases cortisol levels) have given way to models of hormone/behavior interactions and theories that take context into account (*10*). For example, poor attachment, family discord, and low investment in children are believed to affect the timing of puberty onset (*11*). In turn, the combination of these stressors and early puberty contributes to conflict with parents (*12*), lower self-esteem (*13*), and associations with deviant peers (*14*). Neurophysiological and brain imaging studies have demonstrated brain reorganization during adolescence coincident with the onset of puberty, which may make adolescents more sensitive to experiences that affect their judgement (*15*).

### New Designs, New Methods

Time is an important dimension in understanding all developmental stages; at the individual level, in the interaction between individuals and changing sociocultural institutions and practices, and in unfolding historical events. In 1974, Glen Elder published *Children of the Great Depression*, a seminal work based on archival data from the 1920s (*16*). His analysis demonstrated how the growing independence of adolescent boys from their families cushioned the blow of economic shocks on their households. Other longitudinal studies, such as the 1946 British National Birth Cohort Follow-Up Study (*17*), the 1956 New York Longitudinal Study (*18*), and the National Longitudinal Study of Adolescent Health in the USA (www.cpc.unc.edu/addhealth) (*2*), have similarly shown the value of a life-course approach to understanding the development of young people.

**Fig. 1.** Differences in growth and maturation between young women in Bt20, among young women living in the same city and born within 7 weeks of one another in 1990. (The young women and their caregivers gave consent for the photographs to appear in scholarly journals.)

Costly as they are to maintain and complex as the data are to analyze, prospective longitudinal designs are ideally suited to studying human development, especially for understanding processes of change and the temporal ordering of events that are used as one proxy of causality. Follow-up of young people across their life span, as they interact with family, school, peers, and the wider social world, enables understanding that goes much deeper than the snapshot provided by measurement at a single point in time. There are now several large-scale birth cohort studies in developing countries, including Brazil and South Africa, revealing information about long-term and generational effects of nutrition and family life, considered against the backdrop of the demographic and health transitions underway in these countries (*19*, *20*).

The Birth to Twenty (Bt20) study in South Africa enrolled a cohort of more than 3000 children in Soweto-Johannesburg in early 1990. Nicknamed "Mandela's Children," this study has collected data from before birth to age 15 (and the study is planned to continue to age 20) among the children (born immediately after Nelson Mandela's release from prison) and among their families. The second generation, children of the cohort, have started to be born, with the youngest mother only 14 years old when she delivered her baby. This group of young people is the first generation of children to live in a democratic South Africa, and the study aims to portray the development of individuals and groups of individuals as they make their life transitions across a particular historical period (*21*).

### What Are We Learning?

A small group of children can be distinguished in the first 2 years of life whose adjustment difficulties, mainly in the form of problems relating to peers, persist for much of childhood. These problems are fairly well predicted by a combination of physiological (low birth weight), social (single parenthood), and economic (poverty) factors (*22*). Father absence is very high in southern Africa, largely because of migrant labor practices (*23*). Single-parent female-headed households are poorer than others; and men who are not married to the mother, either legally or traditionally, at the time of a child's birth, give progressively less financial support and time to their offspring as they grow older (*24*).

Marked differences occur between young people in all domains, most apparently in physical growth, both in childhood and adolescence (Fig. 1). However, it is not only individual differences that can be studied across time but also the patterns produced by the clustering of personal profiles, conditions, and contexts (*25*). Young people with different characteristics of physical growth or event onset such as puberty or sexual activity can be investigated with respect to the antecedents, consequences, and correlates of particular patterns. For example, children who show signs of rapid "catch-up"weight gain during infancy tend to have greater fat mass, poorer glucose tolerance, and increased risk of obesity, diabetes, and cardiovascular disease in later life (*26*). Boys and girls in Bt20 who were in a more advanced stage of puberty at 13 years of age were more likely to be engaging in a variety of activities, such as smoking, experimenting with drugs, and sexual activity, than were their less developed peers.

Like other developmental stages, puberty has considerable individual variability. Pubertal staging is influenced by a number of generational, social, and biological factors. The age at which

young people enter puberty has declined all over the world, largely as a result of improvements in socioeconomic conditions and nutrition (*27*). In South Africa, for example, menarche has decreased by 0.73 years per decade among girls in urban environments, with the last reported mean age being 13.2 years (*28*), still a whole year later than among African-American girls in the United States (*29*). At the same time, pubertal timing and related physical and psychosocial factors are also strong determinants of risks for adult outcomes, including poor sexual and reproductive health (*30*), social problems (*31*), and chronic diseases in later life (*32*). For example, early menarche is associated with initiation of sexual activity, both early and late puberty are coupled with changes in self-esteem among boys and girls, and weight gain among girls in puberty is related to later risk of hypertension and diabetes.

Accurate measurement of pubertal staging in community-based studies in non-Western societies, particularly Africa, has only recently become possible, with careful validation of self-assessment of hair growth and breast and gonad development against established criteria, such as Tanner's Sexual Maturation Scale (*33*). Relating pubertal staging to risk behaviors also requires that the latter be accurately assessed. Questions about behaviors that are unlawful or socially sanctioned (such as sex, drug use, and truancy) have the greatest potential to be underreported. New methods are becoming available to estimate these behaviors and correct these problems. These include the use of biological markers of behavior, such as the detection of salivary cotinine or thiocyanate to determine underreporting of smoking (*34*). Because cotinine, manufactured in the body, is a by-product of nicotine, cotinine measures are a good proxy for ingestion of or exposure to nicotine. Besides cotinine, we are also using urinary leukocyte esterase (ULE) tests in Bt20, in addition to the consistency of reports over time, as a screen for HIV risk and to estimate underreporting of sexual activity among young adolescents (*35*). ULE tests are more sensitive in females than males and require confirmatory microscopy. Furthermore, urinary tract infections can be acquired in ways other than through sexual intercourse. Nonetheless, at age 13 years, positive ULE tests were found in twice the number of girls who reported having been sexually active and in an additional 50% of girls who reported having had sex at 14 and 15 years of age. Positive ULE tests were associated with subsequent information indicating the likelihood of early sexual activity in 16% of adolescents at 13 years of age and in 21% at 14 who reported that they were not sexually active.

Also important are efforts to improve the accuracy of adolescent self-reports of potentially sensitive information. Recent technological developments have given rise to audio computer-assisted self-interviewing (ACASI). Adolescents read questions on the screen or listen to them through headphones in a language of their choice and enter their responses directly through a standard or modified keyboard. Studies in Bt20, Zimbabwe, and Kenya indicate that respondents prefer the privacy afforded, and some young people rate themselves as being more honest in their replies. User problems with the technology are still challenges, awaiting easier-to-use options before its application among adolescents with low levels of education can be expanded (*36*). Other experimental techniques include the presentation of and response to sensitive questions through personal digital assistants and mobile phones.

We are now entering a period where we can build on new capabilities for gathering data. Reliable and valid measurement, especially of data derived from young people's self-reports, is the essential foundation for rigorous evaluation of programs to improve conditions for optimal adolescent development. For the next 5 years of the Bt20 program, we are concentrating on measures of outcomes: achieving in school or dropping out, having an unwanted teenage pregnancy or completing school, weight gain, high diabetes indicators, conflict with the law, and so on. We are also starting to enroll the next generation—young Bt20 boys and girls are starting to have babies—and this gives us an excellent opportunity to study intergenerational advantages and disadvantages.

Bt20 and other longitudinal studies offer insights into predisposing conditions for beneficial and adverse outcomes in the adolescent years. Many of the findings of these studies emphasize the importance of early and systemic intervention. A good start in life, affectionate and stable family relationships, school and neighborhood support for young people's development, and the like, all predict good outcomes for young people in terms of school achievement, adjustment, civic engagement, and future aspirations. However, there are still reasons for optimism even when children are exposed to extremely dysfunctional circumstances (*37*, *38*). For example, three decades ago, Garmezy and his colleagues found that although having a parent with schizophrenia did increase children's risk for the illness, 90% of the chidren they studied had "good peer relations, academic achievement, commitment to education and to purposive life goals, early and successful work histories" [(*37*), p. 114]. This is also true of conditions of poverty, conflict and violence, and parental substance abuse and criminality; most children exposed to these conditions grow up to lead successful lives as adults, with the capacity to love and work (*38*). Self-stabilizing tendencies enable many children and young people growing up in difficult circumstances to take advantage of even slender opportunities to participate in social activities with others, achieve at what they do and be valued, and contribute to the well-being of those they care about. Opportunity niches can be created by winning a race, being selected for a team or cast in a school play, or having a supportive family member even if not a parent, including a teacher who shows interest. All of these can change how a child sees him- or herself and how others see and treat him or her.

We have been surprised by the extent to which this is true in the Bt20 cohort. Of the 2300 children followed up to age 16, over 50% lived in very poor conditions (less than $1 per day per person), 20% frequently went to bed hungry during their early years, more than 40% had direct or vicarious experience of community or family violence, and only two out of five children had ever lived with their fathers (*23*). Despite these conditions, we currently estimate that only about 5% of the children showed persistent behavioral difficulties in their preschool and early school years, had started smoking or carrying a weapon by age 14, or been in conflict with the law. However, as young people enter into their teen years, travel further away from home to school, are subject to less parental monitoring and supervision, and are increasingly exposed to peers who engage in risky behavior behavior, the rates of potentially problematic behaviors increase. For example, although only 1.6% of 13-year-olds (3.3% of boys and 0.8% of girls) have had sexual intercourse, this rate rises to close to 20% at 15 years of age (27% among boys and 12% among girls). A composite risk score, combining rates of smoking, alcohol and drug use, foreplay, and weapon carrying, shows a significant increase with increased pubertal development and the transition from primary to secondary schooling. Young people most likely to be taking risks in their early teens are those who are advanced in their pubertal development for their age and in environments where they are exposed to older adolescents, without monitoring and supervision by caring adults.

## Conclusion

The adolescent years, and especially puberty, link the impact of generational and early-childhood factors to adult outcomes. Longitudinal studies are demonstrating that it is also an age of opportunity: Good nutrition and healthy lifestyle, positive family and school influences, and access to supportive services, among other factors, can help young people break early patterns leading to ill health and poor social adjustment, with benefits for adult well-being and the next generation of children and youth. New methods are increasing the accuracy and validity of data collected from young people, and developments in both the biological and social sciences are providing unprecedented opportunities (*39*). However, we still know a lot more

about what goes wrong in adolescence and why, and a lot less about how to prevent problems and how to get young people back on track, especially in those areas of the world in which young people face the greatest challenges. New knowledge is being driven by the need to develop and test interventions to promote the physical and psychological well-being of young people and counteract the risks associated with this developmental stage. This is especially true in a world in which many adolescents face the same threats—incomplete or poor-quality education, limited prospects for satisfying work, marginalization, HIV/AIDS and other sexually transmitted infections, substance abuse, violence, anxiety and depression—without the same opportunities for help and support.

**References**


1. See http://esa.un.org/unup/.
2. R. Rindfuss, *Demography* **28**, 493 (1991).
3. J. Kett, *Rites of Passage: Adolescence in America, 1790 to the Present* (Basic Books, New York, 1977).
4. B. Brown, W. Larson, T. Saraswathi, *The World's Youth—Adolescence in Eight Regions of the Globe* (Cambridge Univ. Press, Cambridge, 2002).
5. P. Dasen, *Int. J. Group Tensions* **29**, 17 (2000).
6. A. Schlegel, H. Barry, *Adolescence: An Anthropological Enquiry* (Free Press, New York, 1991).
7. E. Fuchs, *Youth in a Changing World* (Free Press, New York, 1976).
8. C. Delaney, *Adolescence* **30**, 891 (1995).
9. J. Hoover, *Reaching Today's Youth* **3**, 2 (1998).
10. E. Susman, *J. Res. Adolescence* **7**, 283 (1997).
11. J. Belsky, L. Steinberg, P. Draper, *Child Dev.* **62**, 647 (1991).
12. L. Steinberg, J. Hill, *Dev. Psychol.* **14**, 683 (1978).
13. J. Williams, C. Currie, *J. Early Adolescence* **20**, 129 (2000).
14. D. Haynie, *Social Forces* **82**, 355 (2003).
15. S.-J. Blakemore, S. Choudhury, *J. Child Psychol. Psychiatry* **47**, 296 (2006).
16. G. Elder, *Children of the Great Depression: Social Change in Life Experience* (Univ. of Chicago Press, Chicago, 1974).
17. M. Wadsworth, *The Imprint of Time* (Oxford Univ. Press, Oxford, 1991).
18. A. Thomas, S. Chess, H. Birch, *Temperament and Behaviour Disorders in Children* (New York Univ. Press, New York, 1968).
19. C. Victora, F. Barros, *Int. J. Epidemiol.* **35**, 237 (2006).
20. L. Richter, S. Norris, T. De Wet, *Paediatric Perinatal Epidemiol.* **18**, 572 (2004).
21. O. Barbarin, L. Richter, *Mandela's Children: Growing up in Post-Apartheid South Africa* (Routledge, New York, 2001).
22. L. Richter, R. Griesel, O. Barbarin, in *International Perspectives on Child and Adolescent Mental Health*, N. Singh, J. Leung, A. Singh, Eds. (Elsevier, Amsterdam, 2000), pp. 159–182.
23. L. Richter, R. Morrell, *Baba: Men and Fatherhood in South Africa* (Human Sciences Research Council, Cape Town, South Africa, 2006).
24. L. Richter, *Psychosocial Studies in Birth to Twenty: Focusing on Families* (Birth to Twenty, Johannesburg, South Africa, 2004).
25. N. Galambos, B. Leadbeater, *Int. J. Behav. Dev.* **24**, 289 (2000).
26. N. Crowther, J. Trusler, N. Cameron, N. Toran, I. Gray, *Diabetologia* **43**, 978 (2000).
27. M. E. Herman-Giddens, C. Bourdony, E. Slara, R. Wasserman, *Pediatrics* **107**, 609 (2001).
28. N. Cameron, C. Wright, *S. Afr. Med. J.* **78**, 536 (1990).
29. B. Ellis, *Psychol. Bull.* **130**, 920 (2004).
30. J. Udry, *J. Biol. Sci.* **11**, 411 (1979).
31. X. Ge, R. Conger, G. Elder, *Dev. Psychol.* **37**, 404 (2001).
32. C. Berkey, A. Gardner, G. Colditz, *Am. J. Epidemiol.* **152**, 446 (2000).
33. S. Norris, L. Richter, *J. Res. Adolescence* **15**, 609 (2005).
34. K. Bauman, G. Koch, E. Bryan, N. Haley, M. Downton, M. Orlandi, *Am. J. Epidemiol.* **130**, 327 (1989).
35. J. Marrazzo, C. White, B. Krekeler, C. Celum, W. Laffery, W. Stamm, H. Handsfield, *Ann. Intern. Med.* **127**, 796 (1997).
36. B. Mensch, *Demography* **40**, 247 (2003).
37. N. Garmezy, *Am. J. Orthopsychiatry* **41**, 101 (1971).
38. S. S. Luthar, E. Zigler, *Am. J. Orthopsychiatry* **61**, 6 (1991).
39. E. Susman, *J. Res. Adolescence* **7**, 283 (1997).


10.1126/science.1127489

PERSPECTIVE

# Politics and the Life Cycle

**Donald R. Kinder**

**The study of politics and the life cycle began with a rather single-minded focus on childhood and the family—on the idea, as Tocqueville famously put it, that the entire person could be "seen in the cradle of the child." Politics does begin in childhood, and parents do influence their offspring, but change takes place over the entire span of life. I take up the early emergence of partisanship and essentialism, the formation of generations, politically consequential transitions in adulthood, and the rising of politics and its final decline.**

My assignment is to highlight important findings and promising developments in the study of politics over the life cycle: politics, one could say, from the cradle to the grave. I focus on exemplary cases, confine my attention to the United States, overlook many local skirmishes, and concentrate primarily on tracing the life cycle path of political belief (rather than action). The essay proceeds chronologically, beginning with childhood.

## Partisanship and Essentialism in Childhood

What Freud did to assumptions of childhood sexual innocence, Greenstein (*1*) and other pioneers in the field did to assumptions of political innocence. Children may be naïve and poorly informed when it comes to politics, but they are far from innocent. They express strong attachment to the nation. They think of themselves, proudly, as partisans of one party or the other. They believe that their country and its way of life are best. They happily subscribe to conventional stereotypes about blacks and whites, men and women, and rich and poor (*2*).

As to the origins of such beliefs, speculation centered initially on the family, on the claim from social learning theory that children would imitate and eventually internalize what their parents said and did (*3*). The best evidence comes from the landmark study carried out by Jennings and Niemi in 1965, based on independent interviews with a national sample of high school seniors and their parents. Jennings and Niemi discovered that adolescent children did indeed seem to copy the party identification of their parents (when they could discern it). On most matters of belief, however, correspondence between parents and offspring was unimpressive: moderate on attitudes toward social groups and close to negligible on matters of policy. Weak correspondence presumably reflects a failure to communicate: confusion or conflict among parents, the evasion of political subjects in family discussions, and (perhaps especially) indifference to politics among the children (*4*).

An alternative model for parental influence is genetic transmission. Until quite recently, the assumption that political beliefs are acquired through experience has been taken as an article of faith. Rapid developments in human behavioral genetics have made this stance increasingly difficult to maintain. A number of studies have compared the political views expressed by monozygotic twins (who share an identical genetic inheritance) to the views expressed by dizygotic twins (who develop from two separate eggs fertilized by two separate sperm). The comparisons take into account whether twins are reared together or apart, and whether they are raised by their biological parents or adopted into new families. The results suggest that adult political beliefs—on the death penalty, say, or on school prayer—have a sizable genetic component. On matters of politics, parents may influence their biological offspring as much through the "genetic blueprint" they provide at conception as through the modeling and instruction they supply later on (*5*–*7*).

Children are not only partisans; they are also, according to a recent line of research, essentialists. That is, children seem to believe that certain social categories are "natural kinds": real (not constructed) and discovered (not invented). More specifically, they believe that race and sex and ethnicity

Department of Political Science, University of Michigan, Ann Arbor, MI 48106, USA. E-mail: drkinder@umich.edu

belong entirely to the natural world and that differences between, say, blacks and whites are rooted entirely in biology, or blood, or some such underlying essence. Essence "explains" inner qualities—temperament, intellect, character—as well as outward, physical ones. Children come to believe all this, moreover, without instruction. Essentialism is thought to be the product of evolutionary adaptation: The human cognitive system is predisposed to treat social groupings as natural kinds (*8*, *9*). This line of work is relevant here because it suggests a deep foundation for social stereotyping, and social stereotyping is in turn a pervasive feature of adult political belief. Essentialism conceived of in this way clarifies why so much of public opinion is "group-centric": why views on policy, attachments to party, and votes for candidates depend so decisively on the beliefs and feelings people harbor toward prominent social groupings (*10*).

### Political Generations

Political generations are created out of the conjunction of individual development and political history. The formation of a distinctive generational perspective requires both the openness of late adolescence and early adulthood and the intrusion of events such as war, depression, and social disorder. Under these circumstances, a generation is expected to enter political life with a distinctive and largely permanent commitment to a certain point of view—in possession, Mannheim would say, of its own particular "historical-social consciousness" (*11*).

But as a raw empirical matter, instances of historical-social consciousness are not that easy to find. This is in large part because comprehensive worldviews are exceedingly uncommon among ordinary citizens, whatever their generation (*12*, *13*). If we look for less majestic empirical outcroppings due to generation, however, there is good evidence to be found. People command more vivid memories and deeper knowledge for events that take place during their late adolescence and early adulthood. The Great Depression, World War II, the civil rights movement, the Vietnam War: these events loom especially large in the memory of those Americans who were just coming of age at the time. People know more about such events. They invest them with greater importance. And perhaps most important, they are predisposed to apply the lessons of their generation's experience to contemporary matters (*14*, *15*).

Generations are also centrally implicated in partisan change. Under normal circumstances, party identification is a "durable attachment, not readily disturbed by passing events and personalities" [(*16*), p. 151; (*17*, *18*)]. But periods of serene stability are occasionally interrupted. The exemplary case of party realignment in the United States was set in motion by the economic calamity that overwhelmed the country during the administration of Herbert Hoover. In 1932, at first opportunity, Hoover was driven out of office; shortly thereafter, the Republican Party lost control of Congress; by the end of the Depression, the Democrats had become the majority party. This transformation was accomplished in large part by Franklin Roosevelt and the Democratic Party capturing, more or less permanently, the loyalty of those Americans who were about to enter political life in the 1930s. Older generations born before 1905 were predominantly Republican and remained so, whereas younger generations born after 1916 became, and by and large stayed, Democrats (*16*, *19*).

### Transitions

A generational analysis implies that political development takes place primarily in the impressionable years of late adolescence and early adulthood, and there is some truth to this (*18*, *20*, *21*). But if the pace of political development slows in adulthood, it does not halt.

Passage through adulthood can be analyzed as a series of transitions into (and out of) roles: soldier, parent, neighbor, and more. To some degree, these new roles and the social spaces they occupy—partly individually chosen, partly environmentally supplied—serve as "socialization depots." Faced with new norms and fresh ideas, people change.

An excellent illustration of research along these lines is Huckfeldt and Sprague's (*22*) analysis of the "weak ties" that run through neighborhoods (*23*). In theory, neighborhoods are politically consequential because "they determine proximity and exposure—they serve to structure important elements of involuntary social interaction" [(*22*), p. 36]. In a detailed empirical investigation of an election campaign, Huckfeldt and Sprague show how neighborhoods influence individuals' political beliefs through recurrent processes of social influence.

Another way to conceive of passage through the adult years draws on the idea of self-interest—the seemingly straightforward claim that citizens are predisposed to support parties and policies that advance their own material interests. From this perspective, life cycle transitions become important politically insofar as they generate new incentives and distinctive interests. As people step into new roles—by purchasing a home, enlisting in the armed services, or having children—political beliefs should change accordingly. This seems plausible, but self-interest turns out to be a surprisingly unimportant source of political belief (*24*). Under just the right circumstances—when the material benefits or harms of a proposed policy are substantial, imminent, and well-advertised—self-interest can make a difference, as in the lavishly financed and hotly contested campaign for property tax reform in California in 1978 (*25*). But because these circumstances don't come along very often, self-interest contributes little to an understanding of how political beliefs change over the life cycle.

### Aging

Do individuals grow more conservative as they age? No. Aging is generally unaccompanied by movement to the right (or to the left)—not on particular matters of economics or foreign policy and not on partisanship, either (*26*, *27*).

Intensity of partisanship is another matter, however. Identification with a political party tends to strengthen over the life cycle, evidently

**Fig. 1.** Comstock Images/Corbis

as a consequence of the "sheer temporal accumulation of electoral experience" [(*28*), p. 150; (*18*)]. Electoral experience and age are not identical; the former can be affected by changes in laws and procedures governing who is eligible to vote, the creation or suspension of elections (as in the fascist interlude in 20th-century Germany and Italy), and other external intrusions. The strengthening of partisanship with experience may seem a humble result, but it is not. Strong partisan attachments are vital to the preservation of democratic forms of government, and at the dawn of the 21st century, the world is flush with new and fragile democracies.

### Political Action over the Life Cycle

Belief captures only part of the story of politics and the life cycle. What can be said (if telegraphically) about change and continuity over the life cycle of political action (*29*)?

Participation in politics in the contemporary United States is characterized by huge and persistent inequalities. Those with more income, more skills, and especially more education are much more likely to take part in politics across virtually all forms of participation (*30*). The roots of such inequality are to be found, in part, in the family. Well-educated parents are likely to take part in politics and to create homes in which there is lively political discussion. Children growing up in such environments tend to acquire a set of political predispositions—interest, knowledge, efficacy—that motivate participation in politics later on. Well-educated parents are also likely to have well-educated children, and educational attainment is the single most potent predictor of participation in adulthood. Participation begins in the family, and so too does political inequality (*31*).

Over the past three decades, as inequality has increased, participation in civic and political life has generally declined. It turns out that much of this disengagement from public life is generational. Americans who came of age at the time of the New Deal or World War II entered political life at moments of immense consequence and great common purpose. This particular conjunction of individual development and extraordinary political history appears to have indelibly marked not just memory and belief, as we saw earlier, but also participation in the collective enterprises of politics (*32*).

Transitions play a prominent role in political action as well. Several mechanisms seem to be at work here, but all share the same basic premise: Participation in politics is costly, and potential participants count costs. Participation eats up time and sometimes money; it requires a variety of skills; it entails foregoing other opportunities (*33*).

From this perspective, we would expect that transitions in and out of new roles would depress participation, at least in the short run, and they do. Being laid off, getting married or divorced, having children, or changing residence all result in diminished participation, as time and energy are directed to more pressing personal matters (*30*, *34*).

In the longer run, adult roles are important for political action insofar as they "teach" skills that reduce the costs of participation. Those who speak and write well, or who command the interpersonal arts required for organizing others, are more likely as a consequence to take an active part in politics, and such skills are often developed and refined through involvement in the institutions that constitute civil society: work, neighborhood associations, and religious organizations (*35*).

Transitions are also important from this perspective, as they move people into and out of spaces that are sites for political mobilization. To enhance their chances of winning an election or passing a bill, public officials and political organizations often use various mobilization tactics. They sponsor meetings and rallies, circulate petitions, request contributions, supply citizens with arguments (and even text) with which to bombard their representatives, and more. That is, officials and organizations subsidize the costs of participation that people would otherwise have to put up entirely on their own. When people inhabit roles that make them likely targets of mobilization, they are more likely to take part in politics (*30*).

This framework implies that self-interest might be more important to action than to belief, and this seems to be true as well. When predicting who takes part in politics and who does not, it is useful to know whose interests are directly and immediately at stake (*36*, *37*). Self-interest explains not so much where people stand, but whether they act.

Finally, we know that participation in politics typically begins at a relatively low level in early adulthood, rises steadily through middle age as psychological participation in the wider world broadens, and then declines toward the end of life, as "infirmity defeats experience" [(*30*), p. 141].

### Unfinished Business

The study of politics and the life cycle began with a single-minded focus on childhood and the family. Politics does begin in childhood, and parents do influence their offspring, through instruction and genetic endowment both, but change, we now know, takes place over the entire span of life (*38*). Such change seems to follow two rather distinct paths: one for political belief and another for political action. One challenge for future research is to attend more systematically to the difference between belief and action, and to offer, in the end, an understanding of politics and the life cycle from the perspective of the whole person.

### References and Notes


1. F. I. Greenstein, *Children and Politics* (Yale Univ. Press, New Haven, CT, 1965).
2. For an excellent summary of this research, see D. O. Sears, in *Handbook of Political Science,* F. I. Greenstein, N. W. Polsby, Eds. (Addison-Wesley, Reading, MA, 1975), vol. 2, pp. 96–136.
3. A. Bandura, in *Handbook of Socialization Theory Research*, D. A. Goslin, Ed. (Rand-McNally, Chicago, 1969), pp. 213–262.
4. M. K. Jennings, R. G. Niemi, *The Political Character of Adolescence* (Princeton Univ. Press, Princeton, NJ, 1974).
5. L. J. Eaves, H. J. Eysenck, N. G. Martin, *Genes, Culture and Personality* (Academic Press, London, 1989).
6. J. R. Alford, C. L. Funk, J. R. Hibbing, *Am. Pol. Sci. Rev.* **99**, 153 (2005).
7. For a balanced assessment of much of this literature and an argument for treating children's genetic predispositions and their parents' childrearing regimes as dynamically interrelated, see E. E. Maccoby, *Annu. Rev. Psychol*. **51**, 1 (2000).
8. S. A. Gelman, *The Essential Child* (Oxford Univ. Press, New York, 2003).
9. L. A. Hirschfeld, *Race in the Making: Cognition, Culture, and the Child's Construction of Human Kinds* (MIT Press, Cambridge, MA, 1996).
10. D. R. Kinder, in *Electoral Democracy*, M. B. MacKuen, G. Rabinowitz, Eds. (Univ. of Michigan Press, Ann Arbor, MI, 2003), pp. 13–47.
11. K. Mannheim, in *Essays on the Sociology of Knowledge* (Routledge and Kegan Paul, London, 1952), pp. 276–322.
12. P. E. Converse, in *Ideology and Discontent*, D. E. Apter, Ed. (Free Press, New York, 1964), pp. 206–261.
13. D. R. Kinder, in *Handbook of Social Psychology*, D. Gilbert, S. Fiske, G. Lindsey, Eds. (McGraw-Hill, Boston, ed. 3, 1998), pp. 778–867.
14. D. C. Rubin, T. Rahhal, L. W. Poon, *Mem. Cognit.* **26**, 3 (1998).
15. H. Schuman, C. Rieger, *Am. Soc. Rev.* **57**, 315 (1998).
16. A. Campbell, P. E. Converse, W. E. Miller, D. E. Stokes, *The American Voter* (Wiley, New York, 1960), p. 151.
17. D. Green, P. Bradley, E. Schickler, *Partisan Hearts & Minds: Political Parties and the Social Identities of Voters* (Yale Univ. Press, New Haven, CT, 2002).
18. M. K. Jennings, G. B. Markus, *Am. Pol. Sci. Rev.* **78**, 1000 (1984).
19. For a perceptive critique of the idea of realignment in general, see D. R. Mayhew, *Electoral Realignments* (Yale Univ. Press, New Haven, CT, 2002).
20. M. K. Jennings, L. Stoker, *Aging, Generations, and the Development of Partisan Polarization in the United States* (Univ. of California Press, Santa Barbara, CA, 2005).
21. D. O. Sears, C. L. Funk, *J. Pol.* **61**, 1 (1999).
22. R. Huckfeldt, J. Sprague, *Citizens, Politics, and Social Communication: Information and Influence in an Election Campaign* (Cambridge Univ. Press, Cambridge, 1995).
23. M. Granovetter, *Am. J. Soc.* **78**, 1360 (1973).
24. D. Green, *Am. Pol. Sci. Rev.* **86**, 128 (1992).
25. D. O. Sears, J. Citrin, *Tax Revolt: Something for Nothing in California* (Harvard Univ. Press, Cambridge, MA, 1982).
26. J. A. Davis, *Pub. Opin. Q.* **56**, 261 (1992).
27. P. E. Converse, *The Dynamics of Party Support: Cohort Analyzing Party Identification* (Sage, Beverly Hills, CA, 1976).
28. P. E. Converse, *Comp. Polit. Stud.* **2**, 139 (1969).
29. By political action I mean to imply the full repertoire of options currently available to citizens in democracies for acting together on shared interests: participating in elections, petitioning government, joining voluntary associations, enlisting in social movements, even taking to the streets in protest. See C. Tilly, *The Contentious French* (Harvard Univ. Press, Cambridge, MA, 1986).
30. S. J. Rosenstone, J. M. Hansen, *Mobilization, Participation and Democracy in America* (Macmillan, New York, 1993).

31. S. Verba, K. L. Schlozman, N. Burns, in *The Social Logic of Politics*, A. S. Zuckerman, Ed. (Temple Univ. Press, Philadelphia, 2005), pp. 95–114.
32. R. D. Putnam, *Bowling Alone: The Collapse and Revival of American Community* (Simon & Schuster, New York, 2000).
33. C. Tilly, *From Mobilization to Revolution* (Addison-Wesley, Reading, MA, 1978).
34. L. Stoker, M. K. Jennings, *Am. Pol. Sci. Rev.* **89**, 421 (1995).
35. S. Verba, K. L. Schlozman, H. E. Brady, *Voice and Equality: Civic Voluntarism in American Politics* (Harvard Univ. Press, Cambridge, MA, 1995).
36. M. K. Jennings, *Am. J. Pol. Sci.* **23**, 755 (1979).
37. D. P. Green, J. A. Cowden, *J. Pol.* **54**, 471 (1992).
38. J. Kagan, *Three Seductive Ideas* (Harvard Univ. Press, Cambridge, MA, 1998).
39. I thank T. Brader, L. Stoker, and J. Weiss for excellent advice on an earlier version of this essay.

10.1126/science.1127891

PERSPECTIVE

# Would You Be Happier If You Were Richer? A Focusing Illusion

**Daniel Kahneman,[1] Alan B. Krueger,[1,2]* David Schkade,[3] Norbert Schwarz,[4] Arthur A. Stone[5]**

**The belief that high income is associated with good mood is widespread but mostly illusory. People with above-average income are relatively satisfied with their lives but are barely happier than others in moment-to-moment experience, tend to be more tense, and do not spend more time in particularly enjoyable activities. Moreover, the effect of income on life satisfaction seems to be transient. We argue that people exaggerate the contribution of income to happiness because they focus, in part, on conventional achievements when evaluating their life or the lives of others.**

Most people believe that they would be happier if they were richer, but survey evidence on subjective well-being is largely inconsistent with that belief. Subjective well-being is most commonly measured by asking people, "All things considered, how satisfied are you with your life as a whole these days?" or "Taken all together, would you say that you are very happy, pretty happy, or not too happy?" Such questions elicit a global evaluation of one's life. An alternative method asks people to report their feelings in real time, which yields a measure of experienced affect or happiness. Surveys in many countries conducted over decades indicate that, on average, reported global judgments of life satisfaction or happiness have not changed much over the last four decades, in spite of large increases in real income per capita. Although reported life satisfaction and household income are positively correlated in a cross section of people at a given time, increases in income have been found to have mainly a transitory effect on individuals' reported life satisfaction (*1*–*3*). Moreover, the correlation between income and subjective well-being is weaker when a measure of experienced happiness is used instead of a global measure.

When people consider the impact of any single factor on their well-being—not only income—they are prone to exaggerate its importance. We refer to this tendency as the focusing illusion. Standard survey questions on life satisfaction by which subjective well-being is measured may induce a form of focusing illusion, by drawing people's attention to their relative standing in the distribution of material well-being and other circumstances. More importantly, the focusing illusion may be a source of error in significant decisions that people make (*4*).

Evidence for the focusing illusion comes from diverse lines of research. For example, Strack and colleagues (*5*) reported an experiment in which students were asked: (i) "How happy are you with your life in general?" and (ii) "How many dates did you have last month?" The correlation between the answers to these questions was –0.012 (not statistically different from 0) when they were asked in the preceding order, but the correlation rose to 0.66 when the order was reversed with another sample of students. The dating question evidently caused that aspect of life to become salient and its importance to be exaggerated when the respondents encountered the more general question about their happiness. Similar focusing effects were observed when attention was first called to respondents' marriage (*6*) or health (*7*). One conclusion from this research is that people do not know how happy or satisfied they are with their life in the way they know their height or telephone number. The answers to global life satisfaction questions are constructed only when asked (*8*), and are, therefore, susceptible to the focusing of attention on different aspects of life.

To test the focusing illusion regarding income, we asked a sample of working women to estimate the percentage of time that they had spent in a bad mood in the preceding day. Respondents were also asked to predict the percentage of time that people with pairs of various life circumstances (Table 1), such as high- and low-income, typically spend in a bad mood. Predictions were compared with the actual reports of mood provided by respondents who met the relevant circumstances. The predictions were biased in two respects. First, the prevalence of bad mood was

**Table 1.** The focusing illusion: Exaggerating the effect of various circumstances on well-being. The question posed was "Now we would like to know overall how you felt and what your mood was like yesterday. Thinking only about yesterday, what percentage of the time were you: in a bad mood____%, a little low or irritable____%, in a mildly pleasant mood____%, in a very good mood____%." Bad mood reported here is the sum of the first two response categories. A parallel question was then asked about yesterday at work. Bad mood at work was used for the supervision and fringe benefits comparisons. Data are from (*14*). Reading down the Actual column, sample sizes are 64, 59, 75, 237, 96, 211, 82, 221, respectively; reading down the Predicted column, sample sizes are 83, 83, 84, 84, 83, 85, 85, 87, respectively. Predicted difference was significantly larger than actual difference by a *t* test; see asterisks.

| Variable | Group | Percentage of time in a bad mood | | | |
|---|---|---|---|---|---|
| | | Actual | Predicted | Actual difference | Predicted difference |
| Household income | <$20,000 | 32.0 | 57.7 | 12.2 | 32.0*** |
| | >$100,000 | 19.8 | 25.7 | | |
| Woman over 40 years old | Alone | 21.4 | 41.1 | –1.7 | 13.2*** |
| | Married | 23.1 | 27.9 | | |
| Supervision at work | Definitely close | 36.5 | 64.3 | 17.4 | 42.1*** |
| | Definitely not close | 19.1 | 22.3 | | |
| Fringe benefits | No health insurance | 26.6 | 49.7 | 4.5 | 30.5*** |
| | Excellent benefits | 22.2 | 19.2 | | |

***$P < 0.001$.

[1]Princeton University, Princeton, NJ 08544, USA. [2]National Bureau of Economic Research, Cambridge, MA 02138, USA. [3]Rady School of Management, University of California, San Diego, San Diego, CA 92093, USA. [4]Department of Psychology, University of Michigan, Ann Arbor, MI 48106, USA. [5]Stony Brook University, Stony Brook, NY, 11794, USA.

*To whom correspondence should be addressed. E-mail: akrueger@princeton.edu

generally overestimated. Second, consistent with the focusing illusion, the predicted prevalence of a bad mood for people with undesirable circumstances was grossly exaggerated.

The focusing illusion helps explain why the results of well-being research are often counterintuitive. The false intuitions likely arise from a failure to recognize that people do not continuously think about their circumstances, whether positive or negative. Schkade and Kahneman (*9*) noted that, "Nothing in life is quite as important as you think it is while you are thinking about it." Individuals who have recently experienced a significant life change (e.g., becoming disabled, winning a lottery, or getting married) surely think of their new circumstances many times each day, but the allocation of attention eventually changes, so that they spend most of their time attending to and drawing pleasure or displeasure from experiences such as having breakfast or watching television (*10*). However, they are likely to be reminded of their status when prompted to answer a global judgment question such as, "How satisfied are you with your life these days?"

The correlation between household income and reported general life satisfaction on a numeric scale (i.e., global happiness as distinct from experienced happiness over time) in U.S. samples typically ranges from 0.15 to 0.30 (*11*). The relation between global happiness and income for 2004 with data from the General Social Survey (GSS) is illustrated in Table 2. Those with incomes over $90,000 were nearly twice as likely to report being "very happy" as those with incomes below $20,000, although there is hardly any difference between the highest income group and those in the $50,000 to $89,999 bracket.

There are reasons to believe that the correlation between income and judgments of life satisfaction overstates the effect of income on subjective well-being. First, increases in income have mostly a transitory effect on individuals' reported life satisfaction (*2*, *12*). Second, large increases in income for a given country over time are not associated with increases in average subjective well-being. Easterlin (*1*), for example, found that the fivefold increase in real income in Japan between 1958 and 1987 did not coincide with an increase in the average self-reported happiness level there. Third, although average life satisfaction in countries tends to rise with gross domestic product (GDP) per capita at low levels of income, there is little or no further increase in life satisfaction once GDP per capita exceeds $12,000 (*3*).

Fourth, when subjective well-being is measured from moment to moment—either by querying people in real time with the Ecological Momentary Assessment (EMA) technique (*13*) or by asking them to recall their feelings for each episode of the previous day with the Day Reconstruction Method (DRM) (*14*)—income is more weakly correlated with experienced feelings such as momentary happiness averaged over the course of the day (henceforth called duration-weighted or experienced happiness) than it is with a global judgment of life satisfaction or overall happiness, or with a global report of yesterday's mood (Table 3) (*15*, *16*). This pattern is probably not a result of greater noise in the duration-weighted happiness measure than in life satisfaction (*17*). Other life circumstances, such as marital status, also exhibit a weaker correlation with duration-weighted happiness than with global life satisfaction.

An analysis of EMA data also points to a weak and sometimes perverse relation between experienced affect and income. Specifically, we examined EMA data from the Cornell Work-Site Blood Pressure Study of 374 workers at 10 work sites, who were queried about their intensity of various feelings on a 0 to 3 scale every 25 min or so during an entire workday (*18*). The correlation between personal income and the average happiness rating during the day was just 0.01 ($P$ = 0.84), whereas family income was significantly positively correlated with ratings of angry/hostile ($r$ = 0.14), anxious/tense ($r$ = 0.14), and excited ($r$ = 0.18). Thus, higher income was associated with more intense negative experienced emotions and greater arousal, but not greater experienced happiness.

Why does income have such a weak effect on subjective well-being? There are several explanations, all of which may contribute to varying degrees. First, Duesenberry (*19*), Easterlin (*2*), Frank (*20*), and others have argued that relative income rather than the level of income affects well-being—earning more or less than others looms larger than how much one earns. Indeed, much evidence indicates that rank within the income distribution influences life satisfaction (*21*–*23*). As society grows richer, average rank does not change, so the relative income hypothesis could explain the stability of average subjective well-being despite national income growth. The importance placed on relative income may also account for the stronger correlation between income and global life satisfaction than between income and experienced affect, as life satisfaction questions probably evoke a reflection on relative status that is not present in moment-to-moment ratings of affect. The relative income hypothesis cannot by itself explain why a permanent increase in an individual's income has a transitory effect on her well-being, as relative standing would increase. However, the increase in relative standing can be offset by changes in the reference group: After a promotion, the new peers increasingly serve as a reference point, making the improvement relative to one's previous peers less influential (*24*).

Second, Easterlin (*1*, *2*) argues that individuals adapt to material goods, and Scitovsky (*25*) argues that material goods yield little joy for

**Table 2.** Distribution of self-reported global happiness by family income, 2004. The GSS question posed was "Taken all together, how would you say things are these days—would you say that you are very happy, pretty happy, or not too happy?" Sample size was 1173 individuals.

| Response | Percentage indicating global happiness at family income of | | | |
|---|---|---|---|---|
| | Under $20,000 | $20,000–$49,999 | $50,000–$89,999 | $90,000 and over |
| Not too happy | 17.2 | 13.0 | 7.7 | 5.3 |
| Pretty happy | 60.5 | 56.8 | 50.3 | 51.8 |
| Very happy | 22.2 | 30.2 | 41.9 | 42.9 |

**Table 3.** Correlations between selected life circumstances and subjective well-being measures. The question posed was "We would like to know how you feel and what mood you are in when you are at home. When you are at home, what percentage of the time are you in a bad mood____%, a little low or irritable____%, in a mildly pleasant mood____%, in a very good mood____%." The last two response categories were added together to obtain the percentage of time in a good mood. Duration-weighted "happy" is the average of each person's duration-weighted average rating of the feeling happy over episodes of the day, where 0 refers to "not at all" and 6 refers to "very much," and each individual's responses were weighted by the duration of the episode. Sample consists of 740 women from Columbus, Ohio, who completed the DRM in May 2005 (*16*).

| Characteristic | Life satisfaction | Amount of day in good mood (%) | Duration- weighted "happy" |
|---|---|---|---|
| Household income | 0.32*** | 0.20*** | 0.06 |
| Married | 0.21*** | 0.15*** | 0.03 |
| Years of education | 0.16*** | 0.13*** | 0.03 |
| Employed | 0.14*** | 0.12** | 0.01 |
| Body mass index | −0.13*** | −0.08* | −0.06 |

*$P$ < 0.05; **$P$ < 0.01; ***$P$ < 0.001.

**Table 4.** How is time spent and do the activities bring happiness? Time allocation is weighted-average percentage of the nonsleep day for each sampled observation from the American Time-Use Survey (*30*). Weighted average of weekday (5 out of 7) and weekend (2 out of 7) is presented. Sample consists of 3917 men and 4944 women age 18 to 60. Last two rows were computed by authors from a DRM survey of 810 women in Columbus, Ohio, in May 2005; if multiple activities were performed during an episode, the activity refers to the one that was selected as "most important."

| Family income/Gender | Active leisure | Eating | Passive leisure | Compulsory | Work and commute | Other |
|---|---|---|---|---|---|---|
| Men | | | *Time allocation (%)* | | | |
| <$20,000 | 6.6 | 6.6 | 34.7 | 20.8 | 29.1 | 2.1 |
| $20,000–$99,999 | 8.1 | 7.2 | 26.4 | 21.8 | 35.4 | 1.1 |
| $100,000+ | 10.2 | 8.6 | 19.9 | 23.6 | 36.9 | 0.8 |
| Women | | | | | | |
| <$20,000 | 5.3 | 5.7 | 33.5 | 35.6 | 18.5 | 1.4 |
| $20,000–$99,999 | 7.5 | 6.7 | 23.8 | 34.3 | 26.7 | 1.0 |
| $100,000+ | 9.1 | 7.0 | 19.6 | 35.9 | 27.3 | 1.1 |
| Women | | | *Feelings (0-6 scale)* | | | |
| Happy | 4.67 | 4.45 | 4.21 | 4.04 | 3.94 | 4.25 |
| Tense/Stressed | 0.92 | 1.17 | 1.30 | 1.80 | 2.00 | 1.61 |

most individuals. Thus, increases in income, which are expected to raise well-being by raising consumption opportunities, may in fact have little lasting effect because of hedonic adaptation or because the consumption of material goods has little effect on well-being above a certain level of consumption (*26*). Moreover, people's aspirations adapt to their possibilities and the income that people say they need to get along rises with income, both in a cross section and over time (*27*).

Finally, we would propose another explanation: As income rises, people's time use does not appear to shift toward activities that are associated with improved affect. Subjective well-being is connected to how people spend their time. In a representative, nationwide sample, people with greater income tend to devote relatively more of their time to work, compulsory nonwork activities (such as shopping and childcare), and active leisure (such as exercise) and less of their time to passive leisure activities (such as watching TV) (Table 4). The activities that higher-income individuals spend relatively more of their time engaged in are associated with no greater happiness, on average, but with slightly higher tension and stress. The latter finding might help explain why income is more highly correlated with general life satisfaction than with experienced happiness, as tension and stress may accompany goal attainment, which in turn contributes to judgments of life satisfaction more than it does to experienced happiness.

The results in Table 4 also highlight the possible role of the focusing illusion. When someone reflects on how additional income would change subjective well-being, they are probably tempted to think about spending more time in leisurely pursuits such as watching a large-screen plasma TV or playing golf, but in reality they should think of spending a lot more time working and commuting and a lot less time engaged in passive leisure (and perhaps a bit more golf). By itself, this shift in time use is unlikely to lead to much increase in experienced happiness, although it could increase tension and one's sense of accomplishment and satisfaction.

Despite the weak relation between income and global life satisfaction or experienced happiness, many people are highly motivated to increase their income. In some cases, this focusing illusion may lead to a misallocation of time, from accepting lengthy commutes (which are among the worst moments of the day) to sacrificing time spent socializing (which are among the best moments of the day) (*28*, *29*). An emphasis on the role of attention helps to explain both why many people seek high income—because their predictions exaggerate the increase in happiness due to the focusing illusion—and why the long-term effect of income gains become relatively small, because attention eventually shifts to less novel aspects of daily life.

### References and Notes


1. R. Easterlin, *J. Econ. Behav. Organ.* **27**, 35 (1995).
2. R. Easterlin, "Building a better theory of well-being," Discussion Paper No. 742, IZA, Bonn, Germany, 2003.
3. R. Layard, *Happiness: Lessons from a New Science* (Penguin Press, London, 2005).
4. D. Gilbert, *Stumbling on Happiness* (Knopf, New York, 2006).
5. F. Strack, L. Martin, N. Schwarz, *Eur. J. Soc. Psychol.* **18**, 429 (1988).
6. N. Schwarz, F. Strack, H. Mai, *Public Opin. Q.* **55**, 3 (1991).
7. D. Smith, N. Schwarz, T. Roberts, P. Ubel, *Qual. Res.* **15**, 621 (2006).
8. N. Schwarz, F. Strack, in *Well-Being: The Foundations of Hedonic Psychology*, D. Kahneman, E. Diener, N. Schwarz, Eds. (Russell Sage Foundation, New York, 1999), pp. 61–84.
9. D. Schkade, D. Kahneman, *Psychol. Sci.* **9**, 340 (1998).
10. D. Kahneman, R. H. Thaler, *J. Econ. Perspect.* **20** (1), 221 (2006).
11. E. Diener, R. Biswas-Diener, *Soc. Indic. Res.* **57**, 119 (2002).
12. B. Frey, A. Stutzer, *Happiness and Economics: How the Economy and Institutions Affect Well-Being* (Princeton Univ. Press, Princeton, NJ, 2002).
13. A. Stone, S. Shiffman, *Ann. Behav. Med.* **16**, 199 (1994).
14. D. Kahneman, A. Krueger, D. Schkade, N. Schwarz, A. Stone, *Science* **306**, 1776 (2004).
15. In general, we find that the retrospective report of mood on the previous day, which is a global evaluation, shares variance both with the global measures of life satisfaction and with disaggregated measures of emotional experience at particular times.
16. D. Kahneman, D. Schkade, C. Fischler, A. Krueger, A. Krilla, "A study of well-being in two cities," Discussion Paper No. 53, Center for Health and Wellbeing, Princeton, NJ, 2006.
17. We conducted a reliability study of the DRM that asked the same questions of 229 women two weeks apart, and found about the same two-week serial correlation in duration-weighted happiness as in life satisfaction for the respondents.
18. P. Schnall, J. Schwartz, P. Landsbergis, K. Warren, T. Pickering, *Psychosom. Med.* **60**, 697 (1998).
19. J. Duesenberry, *Income, Saving, and the Theory of Consumer Behavior* (Harvard Univ. Press, Cambridge, MA, 1949).
20. R. Frank, *Luxury Fever* (Princeton Univ. Press, Princeton, NJ, 1999).
21. A. Clark, A. Oswald, *J. Public Econ.* **61**, 359 (1996).
22. A. Ferrer-i-Carbonell, *J. Public Econ.* **89**, 997 (2005).
23. E. Luttmer, *Q. J. Econ.* **120**, 963 (2005).
24. W. Runciman, *Relative Deprivation and Social Justice* (Univ. of California Press, Berkeley, CA, 1966).
25. T. Scitovsky, *The Joyless Economy* (Oxford Univ. Press, Oxford, 1976).
26. S. Frederick, G. Loewenstein, in *Well-Being: The Foundations of Hedonic Psychology*, D. Kahneman, E. Diener, N. Schwarz, Eds. (Russell Sage Foundation, New York, 1999), pp. 302–329.
27. B. Van Praag, P. Frijter, in *Well-Being: The Foundations of Hedonic Psychology*, D. Kahneman, E. Diener, N. Schwarz, Eds. (Russell Sage Foundation, New York, 1999), pp. 413–433.
28. See (*31*) for evidence on the misallocation of commuting time and (*14*) on the hedonic experience of commuting and socializing.
29. It goes without saying that happiness is not the only measure of human welfare. Moreover, although income gains may not contribute very much to experienced happiness or life satisfaction, wealthier societies may well enjoy better health care, safer and cleaner environments, cultural benefits and other amenities that improve the quality of life.
30. "Time-use survey—First results announced by Bureau of Labor Statistics," U.S. Department of Labor, USDL 04-1797 (http://www.bls.gov/).
31. A. Stutzer, B. Frey, "Stress that doesn't pay: The commuting paradox," Discussion Paper No. 127, IZA, Bonn, Germany, 2004.


32. The authors thank M. Connolly, M. Fifer, and A. Krilla for research assistance, and the Hewlett Foundation, the National Institute on Aging, and Princeton University's Woodrow Wilson School and Center for Economic Policy Studies for financial support.

10.1126/science.1129688

PERSPECTIVE

# Redistributing Work in Aging Europe

James W. Vaupel and Elke Loichinger

As Europe ages, the proportion of people who work will decline unless older individuals remain in the labor force. Such reform could be part of a more general redistribution of work. If a greater share of the population worked, then the average number of hours worked per week could be reduced. This could particularly help younger people and increase Europe's low birth rates. The challenges facing Germany, Europe's most populous country, are highlighted, but statistics are also given for five other European countries and, for comparison, the United States. Social science research is needed to provide policy-relevant knowledge about life-course options.

Europe, the oldest continent, is growing older. Low birth rates (*1*, *2*), rising life expectancy (*1*, *3*–*5*), and, to a lesser extent, migration flows (*1*) are reshaping the "pyramids" that describe population composition by age and sex (Fig. 1). The new demography poses challenges to current labor practices and policies and offers individuals opportunities for greater life-course choice. We summarize the social science research needed to help societies meet the challenges and to help individuals take advantage of the opportunities. We highlight Germany, Europe's most populous country, but also present data on five other members of the European Union and, for comparison, the United States.

## Two Indicators of Demographic Change

Traditionally, the burden of old-age dependency has been summarized by the ratio of the number of people above 60 to the number of people between 20 and 60. We introduce two "Rostock indicators" that we believe are more informative for summarizing the magnitude of the economic and social challenges caused by population aging (Table 1). The first is based on a simple measure of labor force participation, namely the ratio of nonworkers to workers, with workers including everyone who works for remuneration for at least 1 hour per week (*6*). In 2005 in Germany, this dependency ratio was 1.27: There were five people who were not working for every four people who were (*7*). To determine the impact of demographic change, we can calculate the ratio keeping labor-force participation rates by age and sex at 2005 levels but using the population pyramid for 2025. This dependency ratio for Germany would then be 1.47, with nearly three nonworkers for every two workers. Other European countries show essentially the same picture, but in the United States it improves. The United States faces a less daunting demographic future because women (and men) in the United States are having about two children on average (compared with 1⅓ to 1½ in much of Europe) because of sizable immigration flows of young workers, and because life expectancy has been relatively low and increasing relatively slowly.

Our second indicator of demographic change is based on the number of hours worked per week per capita. Germans in 2005 worked an average of 16.3 hours per week (*7*). This value is so low because only 44% of Germans worked at all. Demographic change from 2005 to 2025 will result in an 8% decrease. In France, Italy, and the Netherlands, population aging will reduce the hours worked per week per capita by about 10%. To a rough first approximation, their economies will be smaller by 10% than they otherwise would have been. If productivity gains are large enough, Europeans may enjoy a somewhat higher standard of living 20 years from now even though they are working less. The distribution of work, however, will be even more unequal than it is today. People will be working less on average because more people will not be working at all.

## Working at Older Ages

To keep dependency ratios and hours worked per week per capita at current levels, it is necessary for age-specific patterns of work to change. Consider Germany. The hours worked per week per capita in Germany in 2005 can be broken down by age (blue line in Fig. 2). If average effort is to be maintained at its current value of 16.3 hours per week, one option would be to increase work by people in their 50s and early 60s (red line in Fig. 2). Not everyone at

Rostocker Zentrum for the Study of Demographic Change and Max Planck Institute for Demographic Research, Konrad-Zuse-Strasse 1, D-18057 Rostock, Germany, and Terry Sanford Institute of Public Policy, Duke University, Durham, NC 27708, USA. E-mail: jwv@demogr.mpg.de

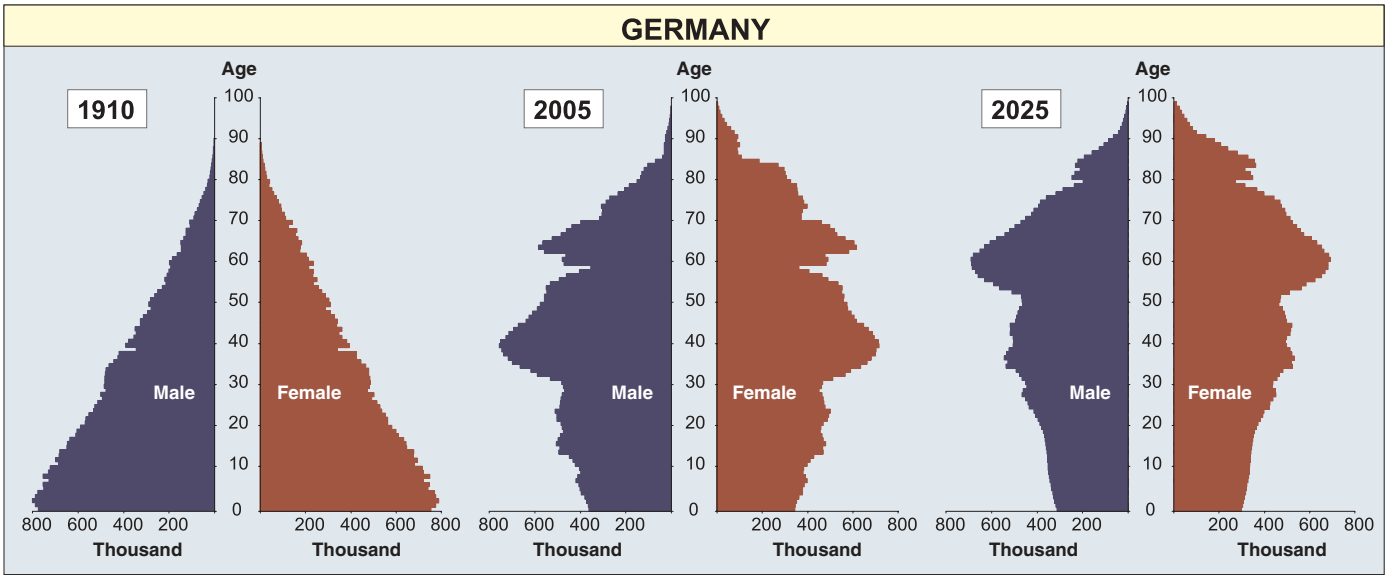



**Fig. 1.** Population pyramids for Germany in 1910, 2005, and 2025 (*7*). The data for 1910 do resemble a pyramid, with many children and few elderly people, but by 2005 there was a bulge of adults around age 40. This bulge will rise to age 60 in 2025.

these ages will be healthy enough to work, but a key finding of recent social science research is that as people live longer they tend to have a longer span of health (*8*). Social science research has also deepened our understanding of the relationship between health and retirement and between health and wealth (*9*). Furthermore, the incentives that drive employee and employer decisions about retirement age are now well understood (*10*, *11*). A knowledge base of demographic and economic theory and evidence exists to inform policymakers and the public about broad needs and options to increase employment at older ages (*12*, *13*).

As the proportion of voters who are older than 50 grows, it may become more difficult to increase the age of retirement. As costs of supporting the elderly rise, expenditures on everything else, including research, education, and child care, may be reduced. This dismal prospect has received much press, but there is little evidence to either support or refute it (*14*, *15*). In the United States and several European countries, intelligent discussion of policy alternatives has created, to varying degrees, a climate of public opinion that recognizes, reluctantly, the need for an increase in the typical age of retirement. In contrast, in France and Italy, public discourse about retirement age (and other economic reforms) is woefully deficient. Social scientists could play a constructive role by participating more actively in public discussions and by putting more emphasis on policy-relevant research.

Some of this research could focus on improving the productivity of older workers through better work environments and lifelong learning. Not everyone has the skills and interests to carry out particular tasks. What kinds of education and organizational arrangements are required to match the labor force with work needs? Many older workers may prefer part-time work. More studies are needed on how to organize 20- and 30-hour work weeks so that they are profitable for organizations and satisfying for individuals.

### Redistribution of Work

If part-time work becomes common for workers above 50 or 60, then more opportunities for part-time work may open up for younger people. As shown by the yellow line in Fig. 2, if people in their 60s and early 70s worked considerably more than today, then work effort could be evenly distributed at a level of about 25 hours per week across ages 20 through 64. This level of effort could be achieved if a few percent were unemployed, a few percent worked 40 hours per week, and the rest worked either 20 or 30 hours per week. The ratio of nonworkers to workers would be cut to a fraction of its current value.

The 20th century was a century of redistribution of income. The 21st century may be a century of redistribution of work. Such redistribution would spread work more evenly across people and over the ages of life. Individuals could combine work, education, leisure, and child-rearing in varying amounts at different ages. This vision is starting to receive some attention from social scientists (*16*–*21*). Achieving it would require radical increases in opportunities to work 20 or 30 hours per week. The Netherlands, Denmark, and Norway may be harbingers of economies with many part-time jobs. Much more research, however, is needed on basic issues concerning the efficiency of such redistribution of work and whether individuals would prefer it.

Future generations may think we (Europeans and Americans) were irrational about the way we spend the time of our lives. We concentrate work in those ages of life when we can have children and when children need the time and energy of their parents. Then, when we are in our late 50s or early 60s, we retire, enjoying decades of leisure, largely paid for by levies on younger adults who are also taking care of children. We concentrate the leisure of our lives in the years when we can no longer have children and when any children we did have no longer need the care they once required.

A redistribution of work might make it easier for younger people to have the number of children they would like to have. The causes of low fertility in Europe, however, are complex and only partially understood (*2*, *22*). Funding for research on policy options has been meager.

How could parents support themselves and their children if they worked only 20 or 30 hours per week? If the need for transfer payments from workers to nonworkers were reduced, taxes and other levies could likewise be reduced. Furthermore, a greater fraction of women, at both younger and older ages, would be in the work force. In principle, it should be possible to redistribute work while maintaining standards of living (*16*). The specifics of how to do this, however, have to be worked out.

Population aging is not going to stop in 2025. Long lives are the probable destiny of most people alive today in developed countries (*1*, *3*–*5*). Extended life spans make life-course flexibility more desirable for individuals and societies (*19*, *20*). Social scientists can develop knowledge about how to move from the stultifying regime in Germany, France, and most of the European Union to societies in which individuals have greater choice about how to spend the time of their lives. To have influence, social scientists will have to augment their fascination with new kinds of data collection and more sophisticated methods of statistical analysis with a deeper concern about making their research more directly relevant to policy issues. To supplement recent decades of micro analysis of individual

**Table 1.** Rostock indicators of demographic structure and change. *R* is the ratio of nonworkers to workers, with workers including everyone who works for remuneration for at least 1 hour per week (*6*). *H* is the number of hours worked per week per capita. The values for 2025 and the relative changes from 2005 to 2025 assume change in the population pyramid but no change in labor-force participation or effort by age and sex. For data sources, see (*7*).

| Country | *R*, nonworkers per worker | | | *H*, hours worked per week per capita | | |
|---|---|---|---|---|---|---|
| | 2005 | 2025 | Change | 2005 | 2025 | Change |
| Germany | 1.27 | 1.47 | 16% | 16.28 | 14.95 | –8% |
| Denmark | 0.97 | 1.12 | 15% | 17.46 | 16.11 | –8% |
| France | 1.43 | 1.69 | 18% | 15.09 | 13.63 | –10% |
| Italy | 1.59 | 1.86 | 17% | 15.19 | 13.48 | –11% |
| Netherlands | 1.01 | 1.20 | 19% | 15.31 | 13.88 | –9% |
| UK | 1.09 | 1.19 | 9% | 17.32 | 16.34 | –6% |
| USA | 1.09 | 0.99 | –9% | 18.71 | 18.29 | –2% |

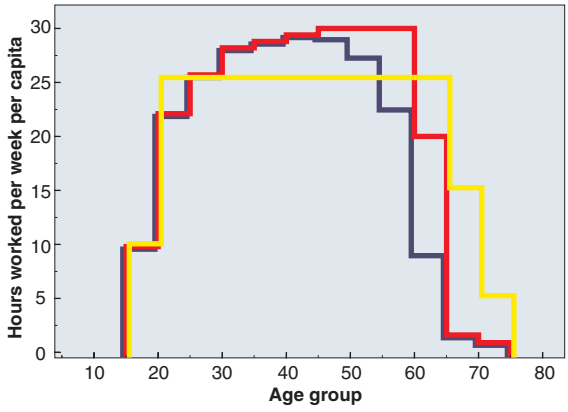



**Fig. 2.** Average hours worked per week by age in Germany (*6*, *7*). The blue line graphs the pattern in 2005 that produced the overall level of 16.3 hours of work per week per capita. The red line shows the increase in work effort by older Germans required to maintain this overall level of effort in 2025. The yellow line illustrates one way to redistribute work more equally while maintaining the overall effort: People between 20 and 65 would work 25.1 hours per week on average.

behavior, researchers need to perform more macro analysis at the population level and more analysis of micro-macro interactions. As discussed by Butz and Torrey in this issue (*23*), such a reorientation could help social scientists contribute more effectively to understanding of demographic challenges and opportunities and a range of other important social issues.

**References and Notes**


1. J. Alho *et al.*, *Changing Population of Europe: Uncertain Future—Final Report* (2005) (available at www.stat.fi/tup/euupe).
2. T. Sobotka, *Postponement of Childbearing and Low Fertility in Europe* (Dutch Univ. Press, Amsterdam, 2004).
3. J. Oeppen, J. W. Vaupel, *Science* **296**, 1029 (2002).
4. J. W. Vaupel *et al.*, *Science* **280**, 855 (1998).
5. K. Andreev, J. W. Vaupel, *Forecasts of Cohort Mortality After Age 50* (2005) (available at www.demogr.mpg.de/papers/working/wp-2006-012.pdf).
6. We used the International Labor Organization's 1-hour-per-week definition of employment (http://laborsta.ilo.org).
7. Data for Fig. 1 are from the German Statistical Office for 1910, from Eurostat for 2005, and from the 10th Coordinated Population Projection, scenario 7, German Statistical Office for 2025. Data for Table 1 and Fig. 2 for Europe are from Eurostat and for the United States from the U.S. Bureau of Labor Statistics. European population forecasts for 2025 in Table 1 and Fig. 2 are averages of forecasts for 2020 and 2030 from (*1*). For the United States the population forecast is the medium UN variant published in 2004. Links to these data sources are as follows: German Statistical Office (www.destatis.de), Eurostat (http://epp.eurostat.cec.eu.int and www.eds-destatis.de), Bureau of Labor Statistics (www.bls.gov), and United Nations (www.un.org/esa/population/publications/WPP2004/wpp2004.htm).
8. V. A. Freedman *et al.*, *Demography* **41**, 417 (2004).
9. L. J. Waite, Ed., *Aging, Health and Public Policy: Demographic and Economic Perspectives* (Population Council, New York, 2004) [*Popul. Dev. Rev.* **30** (suppl.) (2004)].
10. J. Gruber, D. A. Wise, Eds., *Social Security Programs and Retirement Around the World: Micro Estimation* (Univ. of Chicago Press, Chicago, 2004).
11. A. Börsch-Supan, *Labour* **17**, 5 (2003).
12. Organization of Economic Co-Operation and Development, *Reforms for an Ageing Society* (OECD, Paris, 2001).
13. G. Reday-Mulvey, *Working Past 60: Key Policies and Practices in Europe* (Palgrave Macmillan, Basingstoke, UK, 2005).
14. R. H. Binstock, J. Quadagno, in *Handbook of Aging and the Social Sciences*, R. H. Binstock, L. K. George, Eds. (Academic Press, San Diego, CA, 2001), pp. 333–351.
15. M. Blekesaune, J. Quadagno, *Eur. Sociol. Rev.* **19**, 415 (2003).
16. J. Wheelock, J. Vail, Eds., *Work and Idleness: The Political Economy of Full Employment* (Springer, Berlin, 1999).
17. V. W. Marshall, W. R. Heinz, H. Kruger, A. Verma, Eds., *Restructuring Work and the Life Course* (Univ. of Toronto Press, Toronto, 2003).
18. European Foundation for the Improvement of Living and Working Conditions, *A New Organization of Time over Working Life* (Office for Official Publications of the European Communities, Luxembourg, 2003).
19. J. W. Vaupel, *Frankfurter Allgemeine Zeitung*, 8 April, p. 41 (2004). (English translation available at http://user.demogr.mpg.de/jwv/pdf/faz_20040408_41_en.pdf).
20. R. D. Lee, J. R. Goldstein, in *Life Span: Evolutionary, Ecological and Demographic Perspectives,* J. R. Carey, S. Tuljapurkar, Eds. (Population Council, New York, 2004) [*Popul. Dev. Rev.* **30** (suppl.) (2004)], pp. 183–207.
21. D. S. Browning, *From Culture Wars to Common Ground* (Westminster John Knox Press, Louisville, KY, 2000).
22. G. Neyer, G. Andersson, Eds., *Contemporary Research on European Fertility* (*Demographic Research*, special issue 3, 2004) (available at www.demographic-research.org/special/3).
23. W. P. Butz, B. B. Torrey, *Science* **312**, 1898 (2006).


24. We thank W. Butz, P. J. Cook, J. Goldstein, P. Hetze, K. von Kistowski, M. Kreyenfeld, M. Kuhn, H. Kulu, G. Neyer, A. Rasner, S. Schnabel, R. Suzman, and H. Wilkoszewski for helpful comments, and S. Leek and P. Wilhelm for technical assistance. The Rostock indicators of demographic structure and change were developed by the authors at the Rostocker Zentrum for the Study of Demographic Change, a joint venture of the University of Rostock and the Max Planck Institute for Demographic Research. Supported by the Max Planck Society, the Duke University Population Research Institute, and NIH grant AG-08761.

10.1126/science.1127487


PERSPECTIVE

# The Influence of a Sense of Time on Human Development

**Laura L. Carstensen**

**The subjective sense of future time plays an essential role in human motivation. Gradually, time left becomes a better predictor than chronological age for a range of cognitive, emotional, and motivational variables. Socioemotional selectivity theory maintains that constraints on time horizons shift motivational priorities in such a way that the regulation of emotional states becomes more important than other types of goals. This motivational shift occurs with age but also appears in other contexts (for example, geographical relocations, illnesses, and war) that limit subjective future time.**

Most scientists would agree that the explicit study of time falls in the purview of physics, yet interest in various aspects of time spans the natural and social sciences. Time is an integral part of virtually all psychological phenomena. From the sequencing of rewards involved in operant and classical conditioning to the flow of oxygen in the measurement of brain activation, time is built into most behavioral and psychological processes. Psychological science, however, has focused relatively little on the implications of our ability not only to monitor time but also to appreciate that time eventually runs out. I maintain that the subjective sense of remaining time has profound effects on basic human processes, including motivation, cognition, and emotion.

Although change over time is the basic foundation of developmental psychology, theoretical models of human development focus almost exclusively on the passage of time since birth. In child development, this marker has served scientists well. A substantial literature shows that chronological age is an excellent (albeit imperfect) predictor of cognitive abilities (*1*, *2*), language (*3*), and sensorimotor coordination (*4*). At increasingly older ages, however, chronological age is a poorer predictor. Instead, increased heterogeneity or differentiation within samples is considered to be a cardinal feature of life-span development (*5*). Presumably, this is due primarily to differences in experiences and opportunities that individuals encounter over time. Chronic stress, level of education, close relationships, and social status all place individuals on very different developmental trajectories that affect not only day-to-day functioning but also health and longevity (*6*). Late in life, chronological age continues to provide a rough marker of accumulated life experience, but it loses the precision it holds in youth.

A second index of time becomes salient as people grow older, namely the subjective sense of remaining time until death. Although correlated with chronological age, this subjective sense of time gradually becomes more important than time since birth. Because goal-directed behavior relies inherently on perceived future time, the perception of time is inextricably linked to goal selection and goal pursuit. Socioemotional selectivity theory (SST), a life-span theory of motivation, is grounded fundamentally in the human ability to monitor time, to adjust time horizons with increasing age, and to appreciate that time ultimately runs out (*7*). SST maintains that time horizons play a key role in motivation. Goals, preferences, and even cognitive processes, such as attention and memory, change systematically as time horizons shrink. Because chronological age is correlated with time left in life, systematic associations between age and time horizons appear, but findings from experimental studies show that when time perspective is manipulated or controlled statistically, many age differences disappear. In short, across many dimensions, older and younger people behave remarkably similarly when time horizons are equated.

Events like the attacks on September 11th and the severe acute respiratory syndrome

Stanford University, Stanford, CA 94305–2130, USA. E-mail: llc@psych.stanford.edu

(SARS) epidemic in Hong Kong completely eliminated age differences on some measures of motivation (*8*). Young men who suffered from HIV before effective treatments were available seemed to view their social world in the same way that very old people do (*9*). In all of these cases, the fragility of life was acutely primed. The subjective sense of time left was affected and, in turn, equated age differences in preferences and desires.

SST maintains that two broad categories of goals shift in importance as a function of perceived time—those concerning the acquisition of knowledge and those concerning the regulation of emotion states. When time is perceived as open-ended, goals that become most highly prioritized are most likely to be those that are preparatory, focused on gathering information, on experiencing novelty, and on expanding breadth of knowledge. When time is perceived as constrained, the most salient goals will be those that can be realized in the short-term, sometimes in their very pursuit. Under such conditions, goals tend to emphasize feeling states, particularly regulating emotional states to optimize psychological well-being. SST predicts that people of different ages prioritize different types of goals. As people age and increasingly perceive time as finite, they attach less importance to goals that expand their horizons and greater importance to goals from which they derive emotional meaning. Obviously, younger people sometimes pursue goals related to meaning and older people pursue goals related to knowledge acquisition; the relative importance placed on them, however, changes. Indeed, differences between young and old are most striking when goals compete, such as situations in which expanding horizons also entail unpleasant emotional experiences. According to SST, in such cases younger people are far more likely than older people to pursue their goal despite the negative emotional burden. This theoretical shift has helped to make sense of a number of findings in the literature previously referred to as the "paradox of aging" (*10*). Older people were observed to have smaller social networks, to be drawn less than younger people to novelty, and to reduce their spheres of interest; at the same time, however, they were as happy as (if not happier than) younger people. This makes sense if motivational changes with age lead people to place priority on deepening existing relationships and developing expertise in already satisfying areas of life.

However, according to SST, such differences are not due to "age" but to differences in the perception of future time. There are clear age differences in preferences, and these differences can be eliminated by selectively expanding or constraining time horizons (*11*, *12*). For example, asked to choose among three social partners who represent different types of goals (*13*), the majority of older people reliably choose emotionally close social partners. Yet when asked to make the choice after imagining that they just received a telephone call from their physician who told them about a new medical advance that virtually ensures they will live far longer than expected, older peoples' choices resembled those of younger people (*12*). Similarly, when younger people are asked to imagine that they will soon move to a new geographical location, they "look like" older people: they, too, now choose emotionally close social partners (*11*). Thus, endings need not be related to old age or impending death. They need simply to limit time horizons. Preferences long thought to reflect intractable effects of biological or psychological aging appear fluid and malleable.


**Fig. 1.** An example of one pair of advertisements used to study age differences in preferences and memory for products. In each pair, the advertisements were identical except for the slogan. One slogan was related to gaining knowledge. The second promised an emotionally meaningful reward (*14*).

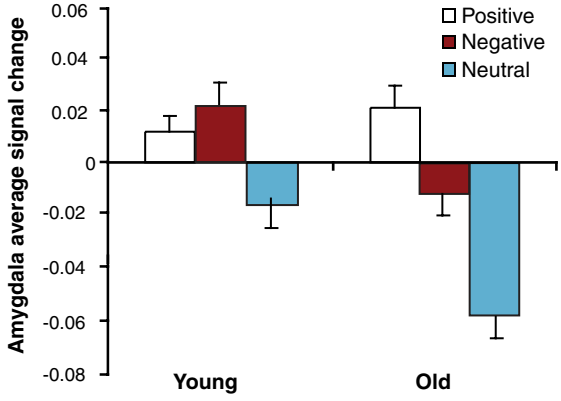



**Fig. 2.** The percentage of signal change in amygdala activation in response to emotionally positive, emotionally neutral, and emotionally negative images. Younger people show significantly increased activation in response to positive and negative images. Older people show increased activation only in response to positive images (*20*). [Adapted from Mather *et al.* (2004)]

We began to explore the ways in which these different motivational states influence information processing. Helene Fung and I developed pairs of advertisements that were identical except for the featured slogan (*14*). In one version of the advertisements, the slogans promised to expand horizons. In the other, the slogans promised more emotional rewards (Fig. 1). The majority of older participants preferred the advertisements featuring the emotion-related slogans. They also remembered these slogans and the products associated with them better than they did the slogans about exploration and knowledge. When older participants were asked to imagine an expanded future before they indicated their preference, they made choices similar to those made by younger participants, that is, they failed to show a significant preference for the emotion-related slogans.

Recently, research has indicated a special preference for emotionally positive information over emotionally negative information in memory in older adults (*15*–*17*). This is particularly intriguing because it has long been known that younger people find negative information more attention-grabbing and memorable than positive information. Indeed, many have posited an evolutionary basis to a preference in memory and attention for negative information. Negative material is richer in information than is

positive material, which often soothes instead of arouses. If the value placed on learning new information changes with shrinking time horizons, however, this preference should dissipate across adulthood. Our research team has coined the term the "positivity effect" to describe a developmental pattern that has emerged in which a selective focus on negative stimuli in youth shifts to a relatively stronger focus on positive information in old age (*16*). Although in some studies, the effect is accounted for primarily by younger people remembering relatively more negative material than positive material, and in other studies the effect is accounted for by older people remembering more positive than negative material, a shift in the ratio of positive to negative across age groups has nevertheless emerged as a reliable finding in the research literature (*18*, *19*).

Of particular interest is recent evidence that older people process negative information less deeply than they do positive information (*20*). While in a brain scanner, older and younger people viewed images of positive, negative, and neutral stimuli. Using event-related functional magnetic resonance imaging, activation in the amygdala was measured in response to the different types of images. Consistent with the results of the behavioral studies noted above, whereas younger adults showed heightened amygdala activation in response to both positive and negative images compared with neutral images, amygdala activation in the older adults increased only in response to the positive images (Fig. 2). Thus, not only at recall but at very early stages of processing, older adults diminish encoding of negative material.

SST suggests that many differences between younger and older people that have long been believed to reflect intractable age differences in attitudes or the consequences of age-related decline may be neither. Young or old, when people perceive time as finite, they attach greater importance to finding emotional meaning and satisfaction from life and invest fewer resources into gathering information and expanding horizons. Tests of hypotheses derived from SST have shed light on the literature showing that, although social networks grow smaller, they also grow more satisfying. Older people appear to prefer such social networks. Hypotheses generated by SST have led to discoveries of differential decline in the processing of certain types of information, suggesting that motivation contributes to at least some observed age differences. As illustrated in the study of advertisement preferences described above, understanding these shifts in motivation can help us to frame information for older adults such that it is more memorable. It also may be that special reliance on emotional responses to options will aid decision-making. Of course, a focus on emotionally satisfying stimuli may be a double-edged sword. Preferential attention to positive information, for example, may contribute to susceptibility to scams or other unscrupulous efforts to take advantage of older people. Many questions remain. It appears, however, that consideration of time horizons can offer insights into the ways in which younger and older people differ, but also show that behavioral differences are often driven by the same underlying mechanisms.

**References and Notes**


1. P. B. Baltes, K. U. Mayer, *The Berlin Aging Study: Aging from 70 to 100* (Cambridge Univ. Press, New York, 2001).
2. T. A. Salthouse, H. P. Davis, *Dev. Rev.* **26**, 31 (2006).
3. D. M. Burke, M. A. Shafto, *Curr. Dir. Psychol. Sci.* **13**, 21 (2004).
4. U. Lindenberger, M. Marsiske, P. B. Baltes, *Psychol. Aging* **15**, 417 (2000).
5. P. B. Baltes, *Dev. Psychol.* **23**, 611 (1987).
6. J. House, *J. Health Soc. Behav.* **43**, 125 (2002).
7. L. L. Carstensen, D. Issacowitz, S. T. Charles, *Am. Psychol.* **54**, 165 (1999).
8. H. H. Fung, L. L. Carstensen, *Soc. Cognit.* **24**, 248 (2006).
9. L. L. Carstensen, B. L. Fredrickson, *Health Psychol.* **17**, 494 (1998).
10. U. Kunzmann, T. Little, J. Smith, *J. Gerontol. B Psychol. Sci. Soc. Sci.* **57**, 484 (2002).
11. B. L. Fredrickson, L. L. Carstensen, *Psychol. Aging* **5**, 335 (1990).
12. H. H. Fung, L. L. Carstensen, A. Lutz, *Psychol. Aging* **14**, 595 (1999).
13. Three prospective social partners are presented: the author of a book you just read, an acquaintance with whom you seem to have much in common, and a member of your immediate family.
14. H. H. Fung, L. L. Carstensen, *J. Pers. Soc. Psychol.* **85**, 163 (2003).
15. S. T. Charles, M. M. Mather, L. L. Carstensen, *J. Exp. Psychol. Gen.* **132**, 310 (2003).
16. M. Mather, L. L. Carstensen, *Trends Cognit. Sci.* **9**, 496 (2005).
17. J. A. Mikels, G. L. Larkin, P. A. Reuter-Lorenz, L. L. Carstensen, *Psychol. Aging* **20**, 542 (2005).
18. S. Schlagman, J. Schulz, J. Kvavilashvili, *Memory* **14**, 161 (2006).
19. D. M. Isaacowitz, H. A. Wadlinger, D. Goren, H. R. Wilson, *Psychol. Aging* **21**, 40 (2006).
20. M. Mather *et al.*, *Psychol. Sci.* **15**, 259 (2004).


21. The research program described herein has been generously supported by grant RO18816 from the National Institute on Aging.

10.1126/science.1127488